

**FUTURE
TRENDS
IN
MICROELECTRONICS
THE ROAD AHEAD**

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

SERGE LURYI

JIMMY XU

ALEX ZASLAVSKY

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE 14 June 1999	3. REPORT TYPE AND DATES COVERED Conference Proceedings	
4. TITLE AND SUBTITLE Future Trends in Microelectronics: The Road Ahead			5. FUNDING NUMBERS F61775-98-WE030	
6. AUTHOR(S) Conference Committee				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of Marseille CRMC2 - CNRS, Campus de Luminy - Case 913 Marseille 13288 France			8. PERFORMING ORGANIZATION REPORT NUMBER N/A	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) EOARD PSC 802 BOX 14 FPO 09499-0200			10. SPONSORING/MONITORING AGENCY REPORT NUMBER CSP 98-1032	
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.			12b. DISTRIBUTION CODE A	
13. ABSTRACT (Maximum 200 words) The Final Proceedings for Future Trends in Microelectronics: Off the Beaten Path, 4 June 1998 - 5 June 1998 This is an interdisciplinary conference. Topics include: microelectronics, nanotechnology, memories, lithography, VLSI, SOI, low power electronics, quantum computing, telecommunications.				
14. SUBJECT TERMS EOARD, C3I, Electronic Devices, Electronics and Electrical Engineering, Microelectronics			15. NUMBER OF PAGES 485	
			16. PRICE CODE N/A	
17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UL	

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

Future Trends in Microelectronics

AQF00-02-0493

19991123 122

Future Trends in Microelectronics The Road Ahead

SERGE LURYI

State University of New York, Stony Brook

JIMMY XU

University of Toronto

ALEX ZASLAVSKY

Brown University



A Wiley-Interscience Publication

JOHN WILEY & SONS, INC.

New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

This text is printed on acid-free paper. ☺

Copyright © 1999 by John Wiley & Sons, Inc.

All rights reserved. Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4744. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158-0012, (212) 850-6011, fax (212) 850-6008, E-Mail: PERMREQ @ WILEY.COM.

For ordering and customer service, call 1-800-CALL-WILEY.

Library of Congress Cataloging in Publication Data:

Future trends in microelectronics / edited by Serge Luryi, Jimmy Xu,
Alex Zaslavsky.

p. cm.

ISBN 0-471-32183-4 (alk. paper)

1. Microelectronics. 2. Nanotechnology. 3. Semiconductors—

Design and construction. I. Luryi, Serge. II. Xu, Jimmy.

III. Zaslavsky, Alexander.

TK7874.F88 1999

621.381—dc21

99-26481

CIP

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

Preface

S. Luryi, J. M. Xu, and A. Zaslavsky

1 FROM MICROELECTRONICS TO NANOELECTRONICS: ROADMAPS AND CHALLENGES

The 1980's and 1990's Microelectronics Logbook: Guidelines for the Future

D. Bois 3

Driving Forces of Future Semiconductor Technology

C. G. Hwang, S. I. Lee, and Y. D. Hong 13

Future Trends in Large-Scale Integrated Circuit Technologies From an Industrial Perspective

Yoichi Unno and Hiroshi Iwai 21

Driving Factors and Breakthroughs for Higher Performance Semiconductor Devices

H. Watanabe 33

On the Edge of Ubiquitous Computing

J. Daniel Janowski and Trey Smith 41

On Future Organization of Hybrid Chip Manufacturing

Eugene A. Feinberg and Serge Luryi 47

The End of Scaling: Disruption from Below

Don Monroe 55

Considerations Beyond Moore's Law

Paul R. Jay 67

Process Technology for Sub-0.1 μm Silicon Devices

Junichi Murota, Takashi Matsuura, and Masao Sakuraba 79

Hot Carrier Degradation Issues in Submicron MOSFETs

K. Hess, L. F. Register, B. Tuttle, J. Lyding, and I. C. Kizilyalli 91

2 BEYOND CMOS: SOI, HETEROSTRUCTURES, THIN FILMS

SOI Technology: Renaissance or Science Fiction? <i>S. Cristoloveanu</i>	105
Single- and Double-Gate SOI MOS Structures for Future ULSI: A Simulation Study <i>Claudio Fiegna, Antonio Abramo, and Enrico Sangiorgi</i>	115
Can Silicon-Based Heterodevices Compete with CMOS for System Solutions? <i>E. Kasper and G. Reitemann</i>	125
Si/SiGeC Heterostructures: A Path Towards High Mobility Channels <i>R. Hartmann, Ulf Gennser, H. Sigg, D. Grützmacher, and G. Dehlinger</i>	133
Potential of SiGe-Channel MOSFETs for a Submicron CMOS Technology <i>J. Alieu, T. Skotnicki, P. Bouillon, J. L. Regolini, A. Souifi, G. Guillot, and G. Brémond</i>	143
Device Implications of Strain Relaxation in Semiconductor Microstructures <i>C. D. Akyüz, H. T. Johnson, A. Zaslavsky, L. B. Freund, and D. A. Syphers</i>	155
The PNP Heterojunction Bipolar Transistor: What Will Be Its Impact in the 21st Century? <i>S. Ekbote, S. Datta, M. Cahay, and K. Roenker</i>	165
Resonance Phase Amplification — A Concept for Operating Si-Based Devices at mm Wave Frequencies? <i>H. Jorke, J. Weller, and J.-F. Luy</i>	173
Silicon Quantum Integrated Circuits <i>D. J. Paul, B. Coonan, G. Redmond, G. M. Crean, B. Holländer, S. Mantl, I. Zozoulenko, K.-F. Berggren, J.-L. Lazzari, F. A. D'Avitaya, and J. Derrien</i>	183
RSFQ Computing: The Quest for Petaflops <i>M. Dorojevets, P. Bunyk, D. Zinoviev, and K. K. Likharev</i>	193
Finite Frequency Shot Noise in Diffusive Wires <i>Yehuda Naveh</i>	207
Short-Channel AIM-SPICE Models for Amorphous Silicon and Polysilicon Thin Film Transistors <i>B. Iñiguez, L. Wang, Z. Xu, T. A. Fjeldly, and M. S. Shur</i>	213

Contents	vii
----------	-----

3 ALTERNATIVE PATHS TO NANOELECTRONICS: SELF-ORGANIZATION, MOLECULAR ENGINEERING

Quantum Dot Lasers: Experimental Results and Future Trends <i>N. N. Ledentsov</i>	223
---	-----

Nanostructure Self-Assembly as an Emerging Technology <i>James L. Merz, Albert-László Barabási, Jacek K. Furdyna, and R. Stanley Williams</i>	237
---	-----

Nonlithographic Fabrication and Collective Behavior for Future Nanoelectronics and Computation <i>A. J. Bennett, D. Levner, J. Li, C. Papadopoulos, A. Rakitin, and J. M. Xu</i>	249
--	-----

Molecular-Scale Electronics <i>Mark Reed</i>	265
--	-----

Organic Molecular Modification of Silicon Surfaces <i>G. P. Lopinski, D. E. Brown, D. J. Moffatt, S. N. Patitsas, D. D. M. Weiner, and R. A. Wolkow</i>	277
---	-----

4 THE MESSAGE IS THE MEDIA: STORAGE MATERIALS AND TECHNOLOGIES

Evolution of Nonvolatile Semiconductor Memory: From Floating-Gate Concept to Single-Electron Memory Cell <i>S. M. Sze</i>	291
---	-----

Double-Junction Gated Single-Electron Transistor EEPROM Cell <i>M. Y. Jeong, Y. H. Jeong, and D. M. Kim</i>	305
---	-----

Single-Electron Memories with Terabit Capacity and Beyond <i>C. Wasshuber, H. Kosina, and S. Selberherr</i>	313
---	-----

New Prospects for Terabit Integration <i>K. K. Likharev</i>	323
---	-----

Data Storage — Is the End of the Bit Near? <i>Arto Nurmikko and Herb Goronkin</i>	339
---	-----

Vertically Integrated SRAM <i>Marco Mastrapasqua, Gerhard Hobler, and Enrico Sangiorgi</i>	353
--	-----

5 ELECTROMAGNETIC SYSTEMS: FROM MICROWAVES TO THE VISIBLE

Electromagnetic Systems Advances

Yoon Soo Park and Max N. Yoder 363

Are Coordinated Roadmaps for Compound Semiconductor-Based Technologies Needed? A Proposal for Smarter Investments

Herbert S. Bennett, Christopher Snowden, and Richard Van Atta 369

21st Century: The Final Frontier for III-Nitrides Materials and Devices

Manijeh Razeghi 381

Properties of III-Nitrides Grown on Si(111) Substrates by Plasma-Assisted Molecular Beam Epitaxy

*M. A. Sánchez-García, E. Calleja, F. B. Naranjo, F. Calle, F. J. Sánchez,
and E. Muñoz* 397

Multi-Wavelength Optical Code Division Multiplexing

C. F. Lam and E. Yablonovitch 407

Photonic Lattices in Semiconductor Waveguides

Jeff F. Young, P. Paddon, V. Pacradouni, T. Tiedje, and S. Johnson 423

Intersubband Terahertz Emitters

Qing Hu, B. Xu, and M. R. Melloch 433

Wide-Bandgap Semiconductor Devices for Future Microwave and Millimeter Wave Power Applications

Wallace T. Anderson 443

Polymer Optical Interconnects

L. Eldada 451

Development of RF-Equivalent Circuit Models from Physics-Based Device Models

S. Luryi 463

LIST OF CONTRIBUTORS 467

INDEX 477

Preface

S. Luryi

Dept. of Electrical and Computer Engineering, SUNY at Stony Brook, Stony Brook, NY 11794-2350, U.S.A

J. M. Xu

Optoelectronics and Emerging Technologies Laboratory, Dept. of Electrical and Computer Engineering, University of Toronto, Toronto, Canada, M5S 3G4

A. Zaslavsky

Division of Engineering, Brown University, Providence, RI 02912, U.S.A.

The celebrated microelectronics industry is rapidly coming to a crossroads. Indeed, for some 50 years, the trillion-dollar industry has been progressing along a one-dimensional path based on miniaturization of integrated circuit components. Further down this path — with the exponential increase in cost (\$2B per IC fab in 1998 and doubling every generation) and the diminishing return (~50% of the revenue goes to cover the fab cost) — all but the biggest players should be squeezed out. As everyone follows the same roadmap, standardization becomes the name of the game. As leaders of innovation become champions of low-cost manufacturing, the innovation also slows down. The much-feared technological maturity is upon us. This plateau should happen well before we run into the so-called fundamental limits, such as quantum fluctuations, *etc.* Taking lessons from the history of science and technology, we can expect that precisely at this point new technologies will emerge and take off.

In the belief that identifying the scenario for the future evolution of microelectronics presents a tremendous opportunity for constructive action today, we have initiated a series of workshops, titled "*Future Trends in Microelectronics*", or FTM. Of course, we are not the only ones concerned about the future of microelectronics, uncertain about further evolution along the beaten path, and interested in alternative directions and new opportunities. There has been no shortage of opinion about what is going on in our profession. Looking down the road, some see storm clouds loom large over major "technological discontinuities". There are pessimists who believe that the microelectronics hardware industry has matured, the research game is over, and the only further progress is in software. Others believe that progress in hardware technology will continue, like it always has, and perhaps even accelerate. Yet many of us look for greener fields outside the fence and for alternatives off the beaten path.

The FTM workshops aim at providing a suitable forum for free-spirited debate among leading professionals in the industry, government agencies, and academia.¹

These meetings convene every three years and are attended by invitation only. Limited by the capacity of the meeting place, over 80 people attended the Ile des Embiez workshop on which this collective treatise is largely based. The workshop format included prepared invited presentations, *ad hoc* contributions and uninhibited exchanges of views and rebuttals, expressions and critiques of new ideas. Some of the key luminaries of our profession shared their opinions and led the discussions on where we are going and/or should be going. Balanced representations of advocacy and opposition were intentionally sought.

To start the debates and help their focus, a number of questions had been raised:

- What is the technical limit to shrinking devices? Does pursuing this limit make economic sense and, if so, in which markets: memories, microprocessors, or both?
- What kind of research does the silicon industry need to continue its expansion?
- What are the anticipated trends in lithography? How does one proceed beyond 130 nm EUV and what are the key limitations: modeling, materials, throughput, cost? Or is it time to consider non-lithographic alternatives, such as self-assembly?
- What can overcome the wiring challenge, beyond the one-shot solution of replacing aluminum wiring with copper? Is there any hope for optical or superconductor wiring?
- Will wide-area electronics be integrated with VLSI? What are the limits to thin film transistors?
- Do we need three-dimensional integration or SOI?
- What are the prospects and constraints of universal wafer bonding?
- To what extent can we trade high speed for low power? Is adiabatic computing in the cards as the ultimate low power electronics?
- What is happening in the evolution (or revolution?) of systems and architectures?
- Where are the big-stake market pulls and pushes for new semiconductor technologies: 3D displays, human-machine interfaces, something altogether new?
- Where is nanoelectronics heading? Given that we can make quantum-effect devices, such as resonant tunneling or single electronics transistors, can we make them broadly useful? Can we get around problems like critical-biasing, wiring, and stochastic fluctuations? Is it reasonable to propose large-scale integration of such devices? Does single-electronics promise ultimate nonvolatile memories?

- Is an architecture revolution (or a second Von Neumann) required?
- Is quantum computing more than a mathematical exercise? Can we have or do we need long-range coherence? What about quantum cellular automata?
- How can we answer the DARPA call for "trillion transistors on a chip"? Challenges and/or fundamental problems?
- Is photonic networking the obvious solution to the telecommunications problem? Zero switching network?
- Low-earth orbit satellite network and satellite-on-wafer, a new cause for rethinking of the fiber-optics system?
- Telecom and computer convergence, implications? Network computers?
- New push for speed and bandwidth from 3D TV, digital photography, and virtual reality?
- Plastic fibers and tipping the balance of GaAs vs. InP?
- Are there green pastures beyond the semiconductor technologies? What can we expect from combinations with superconducting circuits? Molecular devices? Plastic transistors and polymer optoelectronics? Can we hope for bioelectronics with self-produced designed cells? Any prospect for mainstream DNA computing or DNA-assisted self-assembly and packaging?
- Biotech and microelectronics chips combination? Chip-in-brain, is this doable? Acceptable?
- Is there a need for (or the possibility of) integrating compound semiconductor ICs into Si VLSI? What are the merits and prospects of hybrid schemes, such as heteroepitaxy, wafer bonding and packaging?
- What are the most attractive system applications of optoelectronic hybrids? Camcorders? LED-CMOS or LCD-CMOS projection TVs? Large-area imagers and printers?
- What are the possible implications of opto-electro-microwave interactions?
- What can we expect from photonic bandgap structures? Is the "photonic computer" still a realistic hope or a fantasy? Could or should photonics ever replace electronics?
- Progress in wide gap semiconductor technologies, electronic and photonic.
- What is happening in narrow gap semiconductors? What are the current status and prospects of cooling technologies? Are intersubband devices a viable alternative? What are the potential applications of the unipolar intersubband lasers?
- What are current problems and ultimate goals in optical disk memories?

- Changing roles of industrial, government and academic researchers. What is the position of military planners?

The discussion was carried out in a variety of formats. Each of the five workshop days began with presentations by key speakers and concluded with an evening panel session with two or three lead (and provocative) position statements, followed by debates among the panelists (all participants). The debate was forcefully moderated and irrelevant digressions cut off without mercy. Moderators were also assigned the hopeless task of forging a consensus on critical issues. The oral presentations, discussions and debates were complemented by afternoon sessions where the latest experimental results, achievements, and supporting data could be displayed in the form of posters. Formal presentations, each including a twenty-minute question period, comprised the morning session of each day. It was intended and requested that the morning presentations focus not on the description of recent achievements of the speaker or his/her group — these being covered in posters — but concentrate on the future trends in sharp, often provocative, terms. The panel discussions centered about the core issues covered by the morning session and the afternoon poster session of the day.

Compared to the previous FTM meeting (Ile de Bendor, 1995), the Embiez workshop had a much stronger representation from the microelectronics giants of Asia. Perhaps for this reason, or maybe due to rapidly changed times, we saw a greater consensus on several key issues that were expected to be controversial. Many participants, especially those from industry, were quite sanguine about the future until 2015 — expecting to stay on the SIA Roadmap, more or less. Nevertheless, even among the industry types there was a correlation between youth and pessimism (e.g., one younger participant worried about some less capable but cheaper technology eating CMOS for lunch). Perhaps the younger folk view these possibilities soberly because they plan to outlive 2015; then again, perhaps the older optimists have seen so many crises come and go that they cannot be bothered to worry about one more. Significantly, speaker after speaker declared Moore's Law dead, as it applies to the one-dimensional miniaturization trend. Well, ... if Moore's Law is dead, long live Moore's Law, as most agreed that the exponential development of microelectronics would continue with no end in sight! It looks like no one believes that shrinking silicon devices will dominate this exponent any longer. Projections have been made on shifts towards embedded chips and systems in appliances and the human body, the hybridization of electronic, mechanical and biological technologies, *etc.*

There was no shortage of controversial opinion expressed at Embiez, both collectively and individually. As he had done earlier in Bendor, Horst Stormer, the 1998 Nobel Laureate, conducted an opinion poll with questions like "is there any chance that direction XX will bear fruit or is it hopeless?" While it would perhaps be irresponsible to publish the results of this poll at this time, many of the XX practitioners should beware, as the opinions seem to have hardened ...

There has been a discernible shift towards communications: bandwidth, computing in cars and telephones, terabits on the Web, and even microchips implanted in everyone's brain at birth. This shift will undoubtedly have

implications for CMOS, driving the world towards systems-on-a-chip (SOC). But there did not seem to be any clear idea on what these SOC's should do and how technology should get there. Self-assembly was much mentioned, but not yet at a practical level, except perhaps for quantum-dot lasers. Certainly, nothing like intelligent self-replication or learning had been presented in a realistic fashion.

Toward the end of the Workshop it became clear that the attendees collectively have a coherent while multifaceted message that should be shared with the professional community at large. This book represents an attempt to code and transmit this message. It is a collective treatise by all attendees of the Embiez meeting, offered to you thanks to a bold decision by Wiley Interscience, thanks especially to George Telecki, a Senior Editor at Wiley. "Medium is the message" and we are now "in print"!

Acknowledgments

The conference at Ile des Embiez and therefore this book became possible owing to support from:

- National Science Foundation: (DMR, ECS, INT)
- U.S. Department of Defense: (Army, Navy, Air Force, DoD-Europe, DARPA)
- European Union: (Phantoms)
- French Department of Defense: (DRET)
- Industry: (Samsung, Motorola, Toshiba, Nortel, France Telecom)

On behalf of all Workshop attendees, sincere gratitude is expressed to the above organizations for their generous support and especially to the following individuals, whose initiative was indispensable:

- Deborah L. Crawford
- Herbert Goronkin
- LaVerne D. Hess
- Hiroshi Iwai
- Rajinder Khosla
- Jong-Gil Lee
- Yoon-Soo Park
- Gernot S. Pomrenke
- Marc van Rossum
- John M. Santiago, Jr.
- Claudine Simson

- Michael A. Stroschio
- Yoichi Unno
- Albert Zylbersztein

References

1. The first FTM workshop, "*Reflections on the Road to Nanotechnology*", took place in 1995 at Ile de Bendor, France, under the auspices of NATO. Proceedings of the first FTM workshop are available in the book form: S. Luryi, J. M. Xu, and A. Zaslavsky, eds., *Future Trends in Microelectronics: Reflections on the Road to Nanotechnology*, NATO ASI Series E 323, Dordrecht: Kluwer Academic, 1996. The second FTM meeting, "*Off the Beaten Path*", convened last summer (June, 1998) at Ile des Embiez, France.

1 From Microelectronics to Nanoelectronics: Roadmaps and Challenges

The Embiez workshop began with a day of theme setting by representatives of the giant companies that often determine the very pace of the development of microelectronics. As "captains of industry", these individuals were in a good position to offer a clear picture of what matters, why it matters, and what will matter in the future. Their perspectives comprise the main content of Chapter 1.

The critical issues covered in this chapter range from the intricacies of the market growth rates to the device, circuit, and system performance, the integration level and fabrication cost, and to the anticipated technological difficulties and discontinuities. We learn not only about such diverse issues as the predicted lithography bottlenecks, the wafer size and material limitations, the increasing importance of defects and doping variations, the wiring challenge, the new material options and power dissipation, but also about the key new markets and applications, including the industrialists' dreams and fantasies about ubiquitous computing and new manufacturing processes.

Of course, there is no shortage of different opinions. We get views and perspectives from different experts with different backgrounds and interests. But, interestingly, some views are shared and many concerns are common. The wiring challenge and power-dissipation crisis stand out prominently, with no obvious solution in sight. Manufacturing and R&D costs constitute yet another shared concern. Investment levels must keep pace with the cost increases to maintain growth, but it is doubtful that the current high levels of investment can be sustained given exponentially increasing costs that outstrip the industry's market share. So either new markets, new applications, and a new era of microelectronics complexity will relax these constraints, or microelectronics development will inevitably slow down to the much more modest 7% growth rate of the overall electronics industry. The consensus of this chapter's authors is that a new era, described as the microelectronics lateral extension to complexity (RAM-on-processor, system-on-chip, network-on-chip, and superchips) is coming, has to come and should better come and save us!

One interesting observation is that our captains were not at all conservative in their outlook. On the contrary, a critical reader could accuse some of them of wild optimism. It should be noted, however, that every participant in the debates was asked to be critical and provocative and to challenge others. While acknowledging the fact that it is becoming more and more difficult for the industry to accept approaches off the beaten path, our captains displayed no lack of imagination. It turned out that each has his own "pocket map" of the uncharted waters off the beaten path, and we have been offered a chance to have a peek at them. These "maps" contain much information and detail: some are clearly marked and

explained, others open to various interpretations, still others nothing short of inspiring to the perceptive mind.

Curiously, certain issues were notable by their omission. Thus, no one cared to comment on the fact that most of the fundamental technological crises are rooted in the same ground — the von Neumann binary serial computing paradigm. Perhaps, one may move to a new computing paradigm to overcome some of the predicted crises. Perhaps some form of alternative computing, say, massively parallel, non-binary (analog or multi-level) computing will change the entire picture. Hints pointing towards such scenarios are scattered throughout the contributions collected in this and the following chapters.

The 1980's and 1990's Microelectronics Logbook: Guidelines for the Future

D. Bois

France Telecom, CNET, B. P. 98, 38243 Meylan, France

1. Introduction

For the past 50 years, i.e. since the invention of the transistor, the exponential progress of microelectronics has been fuelled by a 15% yearly miniaturization rate, inducing almost 30% cost decrease and 50% performance improvement in all electronic functions each year. In turn, this progress of the technology has been the key driving force for a 15% growth rate of the semiconductor market, which generated enough revenues to cover the exponential increase (also about 15% per year) of the R&D expenses needed to feed such exploding technical progress. The major part of those expenses is nowadays devoted to the increasingly huge manufacturing R&D needed to optimise chip size, wafer size, defectivity, and interconnects — all factors building up to make Moore's Law a reality.

As a guideline for the future of microelectronics, it is worth analysing how the strengths and weaknesses of this Moore's Law engine — shown schematically in Fig. 1 — evolved during the last decades.

2. Major limitations and solutions

The first key event of those years is that, despite doubts and sometimes pessimistic predictions, most limitations identified in the 1980s have indeed been overcome. At that time, because of those limitations, many experts considered that 0.3—0.5 micron would set a limit to miniaturization. Fortunately, they were wrong. Looking at technology changes since then, it turns out that the main limitations have indeed been overcome by a few simple and straightforward technical innovations:

- the self-aligned source-drain structure and the low voltage operation solved all the transistor related issues, namely the parasitic resistance and hot carrier induced instabilities;
- deep UV lasers pushed the lithography well below quarter micron;
- and last but not least, chemical-mechanical polishing allowed both the replacement of the old LOCOS by shallow trenches and a drastic increase of the number of interconnect levels.

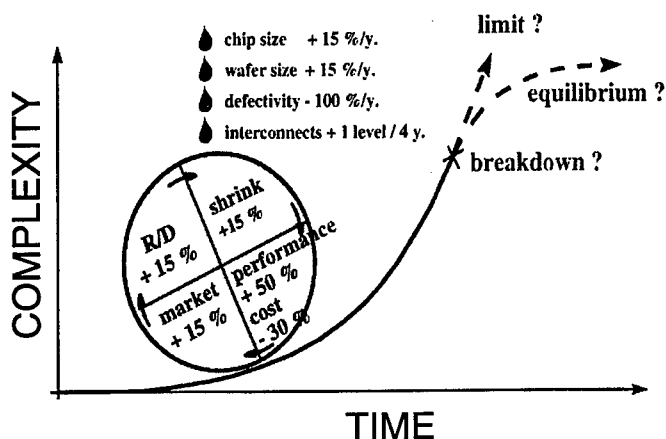


Figure 1. Engine and fuel behind Moore's Law.

On the other hand, all new device architectures, such as vertical MOS, 3D structures, or SOI failed to bring practical solutions to industry. Obviously, this fact indicates that introducing any technical breakthrough is getting more and more difficult, but it also reflects a new phenomenon: the *de facto* standardization of the processes used by the different IC manufacturers for each generation of CMOS technologies. Such convergence results from several factors: decrease in the number of equipment manufacturers for each type of equipment; cell library constraints and second source requirements; and co-operation between competitors. These factors have already had a major effect on industrial R&D, whose objective is no longer to find unique differentiating solutions but to bring standard solutions to the market before competitors. Therefore, "off the beaten path" approaches are accepted by the industry with more and more difficulty.

3. The interconnect crisis

The very bad technical news about this period is that limitations related to interconnects stand up to all innovations; they are just displaced somewhat, getting more and more severe. Indeed, in complex ICs interconnect delays have been overtaking characteristic gate delays, unloaded and loaded, for a long time, but it is only recently that interconnect delays longer than the clock period have been observed — see Fig. 2.

Such delays put severe constraints on circuit design in addition to overall performance limitations. Modelling, simulation, and optimization of interconnects are key challenges for microelectronics R&D; they are mandatory to cope with interconnect parasitic effects, even if they will not make delays smaller. However, these delays will not put a well defined limit to the development of the conventional silicon IC technology: one can survive with large delays, provided adapted architectures and designs are used. For that and other reasons (less power

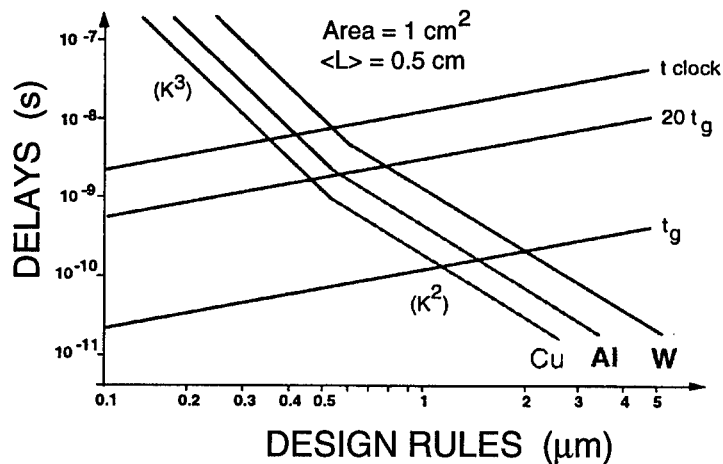


Figure 2. Time delays vs. downscaling of device dimensions.

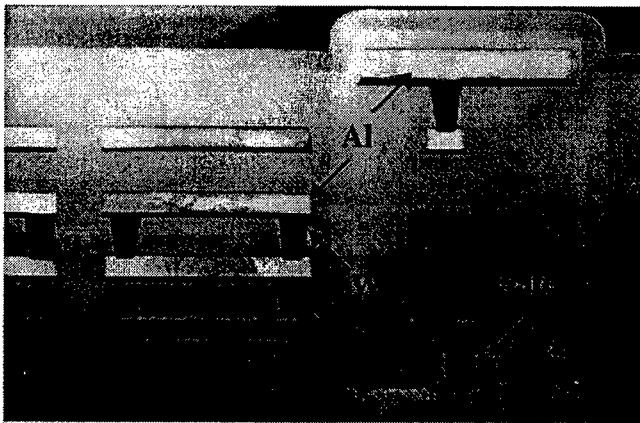


Figure 3. Interconnect levels for 0.35 μm CMOS.

dissipation, higher robustness), asynchronous architectures will be a must in the future. The concept of integrated circuit that has already evolved towards a "system on a chip" will continue its evolution towards a "network on a chip". In digital circuits with several square centimetre areas, the data transmission will possibly make use of networking methodologies, such as the asynchronous time mode (ATM).

Unfortunately, delays are far from being the most severe consequence of the interconnect crisis; undoubtedly, its direct effect on the CV^2 power dissipation is the most important one. In the past, the dynamic power dissipation of ICs has been kept within acceptable limits despite the rapid increase in frequency, because

capacitances were shrunk down together with dimensions. This relationship happened to stop during the last decade. We are no longer playing with the plate capacitor model, but instead with the linear model, which means that the total capacitance in a circuit is just related to the total length of its wires. This length is increasing dramatically, both with the reduction of the wire pitch and with the number of metal levels (see Fig. 3); it will reach several kilometres in 0.1 micron circuits. So the total capacitance related to interconnects dominates over the active device capacitance and thus determines the power dissipation per cm^2 of ICs. Since this power dissipation varies as $\kappa V^2/L$ (where L is the average pitch, V the operating voltage and κ the dielectric permittivity), one has few parameters available to optimise dissipation.

In fact, we are playing our last cards to maintain the power dissipation within acceptable limits. As far as we know, it will be hard to decrease the power supply below 0.7–1 V, and the minimum permittivity is indeed equal to unity, the value for vacuum. Changing the technology to replace SiO_2 by such materials as SiO_F , organic or porous substitutes represents a lot of R&D effort for just a factor of 4 improvement. Remember that microelectronics is usually dealing with orders of magnitude.

The third consequence of the interconnect crisis is that high current transistors are needed to charge the capacitance of the lines within acceptable delays. So, one no longer cares about the intrinsic figure of merit of transistors; it is useless to look for some new low current device until the interconnect crisis is solved. And indeed, the dream of single electron transistors might well fail in front of the interconnect requirements, unless we find some way to operate circuits without charging and discharging kilometres of metal lines to transmit the bits between the various parts of the circuit. That is another challenge for future R&D. A breakthrough is definitively needed in the field of interconnects.

4. The acceleration of technical progress

Despite such critical technical concerns as well as economical issues, surprisingly the progress of silicon ICs has been accelerating during the last decade and it is expected to continue on the same aggressive path over the next one. This acceleration is clearly illustrated by comparing the predictions of the various versions of the famous U.S. Semiconductor Industry Association (SIA) roadmap. Four years ago, quarter micron CMOS was planned to be introduced in 1998, but it was achieved in 1997. Even more impressive is the acceleration of the on-chip clock frequency. Just four years ago, one GHz clock frequencies were expected only by the end of the next decade, and they are already announced for microprocessors.

Such very optimistic views must not hide the fact that Moore's law, behind this silicon roadmap, is not a physical law but the result of a combination of economical and technical forces. One must keep in mind that the exploding progress of microelectronics is a consequence of a far-from-equilibrium situation. Indeed, the magic thing about integrated circuits is that the better they perform, the

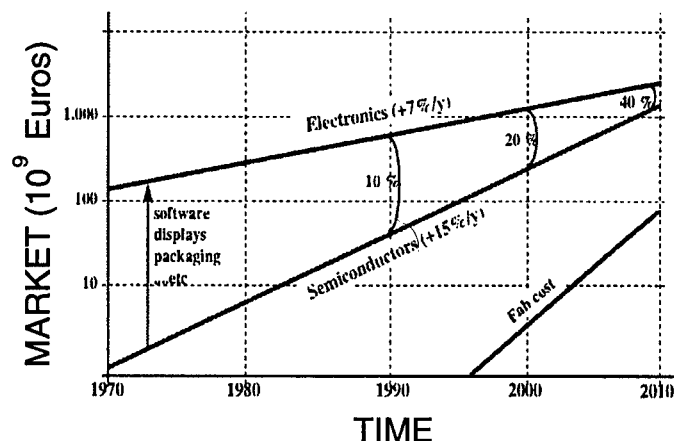


Figure 4. World-wide spending on electronics and semiconductors.

cheaper they are, so until now there has been no technical-economical regulation in this system. The well known and dramatic semiconductor economical cycles are here to recall for us, if necessary, that the counterpart of such an out-of-equilibrium situation is instability and to some extent unpredictability. One crucial question then arises: can the electronic market absorb the semiconductor industry explosion for long, and without any crisis?

5. Towards the end of the semiconductor pervasion phenomenon

As far as the market is concerned, the key event of the last decade is that the semiconductor industry started competing to some extent with its first customer: the electronics industry. Indeed, it has been known for a long time that the semiconductor turnover is growing at a 15% yearly rate, about twice the 7% growth rate of the whole electronics industry. This difference has no consequences until the absolute value can no longer be neglected, which is the case now, since the share of semiconductors in electronics turnover is amounting to almost 20%, and will reach about 40% before the year 2010 by simple extrapolation of the observed exponential growth, as illustrated in Fig. 4. This growth is a major concern because semiconductors are not end products; they cannot be used without surrounding elements, whose costs are at least not decreasing: power supplies, displays, packaging, software etc. A 40% semiconductor-to-electronics ratio is probably the maximum acceptable, and thus the pervasion phenomenon and total semiconductor growth rate might slow down within the next ten years from the historical 15% to a 7% trend.

In the year 2010, the semiconductor turnover will approach 1000 billion Euros, which should imply a silicon consumption per person and per year of about 70 cm², compared to about 12 cm² today. This consumption will correspond to

an annual expense of about 2000 Euros per person, for an average of 1.5 billion consumers, which seems quite acceptable in comparison with other activities, e.g. the automotive sector. But new applications of ICs and new markets for the whole electronics industry must be found to make use of those annual 70 cm², each cm² integrating billions of transistors: it might be one of the key challenges for R&D and marketing people during the coming years. Another way of enlarging the semiconductor market is indeed to increase the consumer number, hopefully to the future 10 billion people world-wide. It is to be noticed that our industry incurs lower environmental costs than any major industry, so one does not see any limitations arising from environmental concerns. That fact might give semiconductors an increasingly important advantage over other industries.

On the other hand, the R&D priority shift from technology to applications that has occurred during the last decade might affect the progress of integration, since both the R&D costs and investments required for the silicon process are still increasing at a high rate. This increase has forced a generalized co-operative approach to processing R&D, through industrial partnerships or national programs. One can now wonder how long this model might account for the increase of R&D costs and whether this co-operative R&D model will be as effective as the competitive one to push ahead the integration progress. One must also remember that most of the process innovations on which we rely nowadays come from 10 years ago, or more, a time when leading companies such as IBM, AT&T and the major Japanese companies were in very different positions from where they are today. Who is now preparing the techniques to be industrialized in the next decade? Who will pay for manufacturing R&D in the future if the competition continues to move towards applications?

6. New market opportunities

Fortunately, the present rapid development of electronic based entertainment, computing and communication tools, especially portable ones, will open large new fields and provide strong technology drivers for microelectronics during the next ten years. The need for low-power high-complexity circuits is a key driving force. But above all, the advanced human interfaces required to make electronic tools acceptable to all consumers, e.g. speech recognition or image processing, will offer new opportunities for digital silicon technologies. At the same time, radio frequency wireless systems will drive the development of high speed analog silicon and possibly silicon-germanium ICs in the multi-GHz range.

Mobile phones and PCs will continue to drive the market for some years but, in view of their fast penetration, both markets will saturate during this period, at least in the most advanced countries; hopefully, they will be replaced by the new wave of networking tools.

The progressive convergence of telecom, computer and entertainment industries, which constitutes an important event of the last period, might indeed modify drastically the whole electronics industry. In particular, the balance between local (individual) data treatment and storage and centralized networked

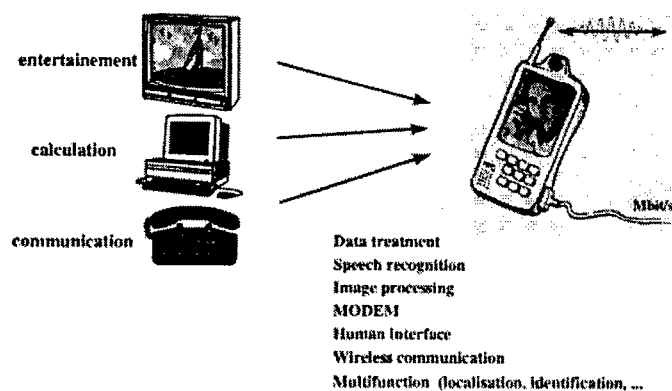


Figure 5. Convergence of computing and telecommunications.

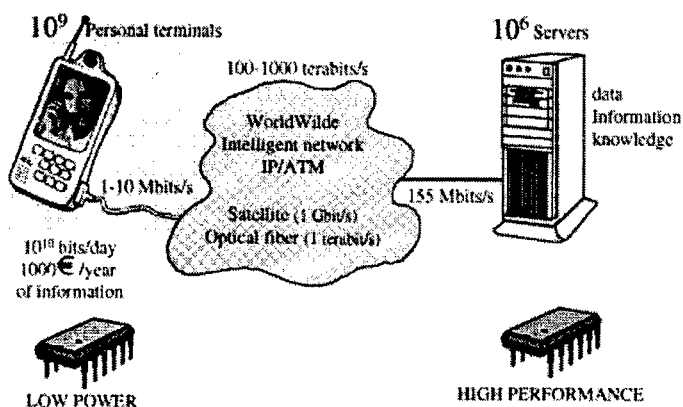


Figure 6. Full network scenario for the years 2010-2020.

tools may change, thus leading to a full set of new products. In a full network scenario, shown in Fig. 6, all the consumer needs for information and entertainment might well be satisfied by a unique multipurpose terminal, possibly portable, making use of a few cm^2 of silicon. Such a terminal would allow people to interact at high speed, typically 1–10 Mbits/s, through a world-wide intelligent network using satellites and optical fibres. A few millions of servers would be enough to satisfy all practical requirements for one billion consumers, through a network with an average data flow rate of a few hundred terabits/s (3 terabits/s are expected within 5 years for the Internet).

Such a world-wide distributed managed information system becomes economically possible since optical communications follow an exponential progress even faster than the silicon one, with a doubling of the transmission rate

each year. This increased bandwidth will make the availability of updated information through the network easier, and maybe cheaper, than locally from a set of CD-ROMs. It might well displace part of the silicon consumption and associated added value from consumers to telecommunication and service providers, which would affect both technical and business requirements. The current discussion about network computers and low cost PCs to provide access to the Internet is a first illustration of such a possible scenario. Should this scenario be confirmed, it would put a limitation to the expansion of the semiconductor market, both in terms of volume and pricing. On the other hand, it would generate new interest in high-performance professional ICs. For instance, such exotic things as low temperature devices could be used in networks: either for the high speed switching and routing or for the huge data treatment inside the servers. Similarly, the base station or satellite requirements for rf communications are very different from the consumer terminal ones. Recently, the first low-temperature analog ICs aimed at such systems have been actually industrialized.

From its origins, the development of ICs has been driven in a very straightforward way by the computer industry, with the DRAM serving as almost the sole technology driver over several decades. The emergence of new diversified markets might stop this simple development paradigm, and make microelectronics enter a more complex era, possibly with different streams more difficult to predict.

7. The industry dreams: a little bit of technology fiction

Since the initial big bang of microelectronics, people have been dreaming of a new technical breakthrough. For many researchers, this dream led to the quest for a new active device to reproduce the fifty year old transistor revolution. The weakness of this approach is in forgetting that the transistor would not have revolutionized electronics without its marriage to planar technology and batch photolithographic processing. So the real dream today for semiconductor manufacturers is to find a new device and associated fabrication methods.

The concept of self-organization that emerged during the last decade is indeed the fruit of such a dream. But one must remember that most of the so-called self-organized high complexity systems in nature grow from seeds. In fact, instead of self-organized, they must be regarded as "self-copying" systems. The true self-organization of complex systems takes millions of years, which is not suitable to satisfy the time-to-market and manufacturing cycle time requirements ... To some extent, seeds act as the masks in our silicon technology: they contain the large amount of information that is necessary to define a complex system. So, if one wants to make this dream a technical and industrial reality some day, one must find the way to replace seeds by some manufactured object. This replacement is physically conceivable. If one accepts a little bit of science fiction, collectively micropatterned substrates might be used as seeds to grow self-organized active elements such as nanostructures or molecules in a well defined array, having the collective behaviour of a well specified complex system. This dream of a self-assembled system is illustrated in Fig. 7.

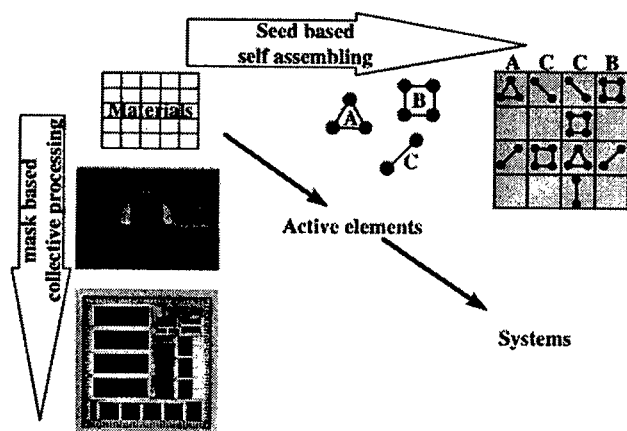


Figure 7. The dream of self-assembled systems.

Another dream for all technologists is to get rid of the defect nightmare. Indeed, since the invention of writing, information has been memorized by patterning symbols into various solids, which implies that any localized defect generates an error, possibly in the whole system behavior. Again, if one accepts some technology fiction, one can imagine a drastic change in the way we have been memorizing the information for several millennia. We know that any time-independent medium can be used to memorise information, not only static ones such as solids, but also dynamic ones such as waves. For instance, an optical fiber is indeed a high-capacity dynamic RAM, with just two active elements at both ends, to write-read and refresh the data. This observation shows that there is no physical direct relationship between the device complexity and the number of memorized bits, which implies that the race for low defectivity, with its impact on manufacturing cost, might change its course in the future.

Those two incursions into science fiction should help us consider that the information technology roadmap, if we look at it on a very large scale, is really widely open. Some new concepts might revolutionize the information technology in the future, as in the last thousands of years did the invention of writing, the collective processing paradigm, optical processing, and at last the transistor.

8. Conclusion

Microelectronics has become a major industry, with still a large technical and economical potential, provided a huge R&D effort can be devoted to manufacturing and applications in the coming years. If so, the performance limitations that we see today will probably be solved, similarly to what happened with the limitations one anticipated in the 80's. But those limitations are no longer related to small

geometries and active devices, but mostly to complexity and defectivity, to interconnects and the related power dissipation. Hence the R&D effort must be somewhat reoriented towards this field. An interdisciplinary "off the beaten path" approach is mandatory to overcome future limitations: the interconnect scheme, the algorithms, the circuit architecture and design must all be considered together to find new ways for digital signal processing without charging and discharging kilometers of wires at high frequency. On the other hand, the industrial interest in research oriented towards new active device concepts of any quantum type will remain limited, unless they address either a large reduction of the operating voltage to overcome the power dissipation issue, or allow a new low cost and robust processing paradigm. This kind of advance is mandatory to break the process complexity and the related defectivity/cost industrial nightmare ; as in the past did the invention of planar collective transistor manufacturing, or even further the invention of printing.

As far as the market is concerned, one can no longer rely on the pervasion phenomenon of the whole electronics industry to broaden the semiconductor market and maintain during the next decades the 15% historical growth rate necessary to reach a 1000 billion Euros turnover before year 2010. The semiconductor industry must generate new applications in order to broaden the electronics market itself; this is the second key challenge for microelectronics R&D. One must find a way to multiply by a factor of six the average yearly silicon consumption per person to reach this goal.

The occurrence of the electronically assisted information society, where entertainment, communication and business will be carried out through the global telecommunication network, will provide numerous market opportunities and new technology drivers. But, it might modify the balance between the development of low cost local systems and centralized high performance ones, thus changing some IC technical requirements. Moreover, since the DRAM is no longer the unique technology driver, and such devices as low power microprocessors, digital signal processors, image processors, rf analog circuits also play an important role, the evolution of the technology might well be more complex, diversified, and maybe more unpredictable than it has been during the past decade. The shift of R&D from technology towards software, application and usage, which has been one of the key turns during the last years, is a major concern for the long term progress of the silicon technology, which still requires exponentially increasing expenses. So, we might enter during the following decade a new era with some stabilization of the mainstream technical progress, leaving the way open to a more diversified development, with possibly new "start-up" technologies. This is an opportunity for "off the beaten path" approaches, provided they address new manufacturing, information processing, or application paradigms.

9. Acknowledgments

The author is grateful to A. Chantre for reviewing the manuscript and G. Auvert and R. Pantel for providing some of the illustrations.

Driving Forces of Future Semiconductor Technology

C. G. Hwang, S. I. Lee, and Y. D. Hong

Semiconductor R&D Center, Samsung Electronics, San #24 Nongseo-Ri, Kiheung-Eup, Yongin-City, Kyungki-Do, Korea

1. Introduction

Since the appearance of the first commercial semiconductor products there has been a tremendous amount of progress in technology, applications, and volume. Today, however, the semiconductor industry is facing technological limitations in mass production and is suffering from a cost and price crisis. The cost of manufacturing facilities has almost doubled from generation to generation, resulting in a dramatic increase of manufacturing and R&D costs. However, the price per transistor has declined at the rate of 21% per year on the average, with the decline accelerating to more than 35% per year over the past five years. While we are making great efforts on new technological concepts, the industry should concentrate on cost effectiveness to continue the semiconductor business and even to survive from the competition.

Since the development of the industry's first microprocessor in 1974, semiconductor logic and memory devices have seen rapid market growth and dramatic development in performance, density, and die size. To date, logic devices with clock frequencies in the GHz range and area in the hundreds of mm² have been developed by overcoming technological limitations. It is expected that the semiconductor marketplace will reach US\$ 250 billion in the year 2000 and expand to US\$ 450 billion by the year 2005. However, further development in performance and density cannot be achieved without significant breakthroughs in technology as well as aggressive investment from chip manufacturers.

First, this article will discuss a possible way of overcoming the currently unbalanced situation between bit cost and price in the semiconductor industry. Second, the technological limitations expected in the next decade will be reviewed, and the driving forces for future semiconductor technology will be introduced, including potential breakthroughs in several key technologies, device architecture, and value-added devices.

2. A possible solution to the bit price crisis

According to Moore's Law, the drop rate of cost per transistor (bit cost) is about 50% every three years. However, the historical cost trend of DRAM pricing shows that the bit cost has been decreasing 24% every year from generation to

generation, but decreasing 45% every year within the same generation. The difference is mainly due to somewhat slower cost reduction beyond the fourth year after the introduction of each generation. Recently, the drop rate of the bit cost in memory chips has been more precipitous than that of the bit price. For DRAM products, the drop rate of the bit price between 1996 and 1997 was about 40% per year, while that from the end of 1980s to 1995 was about 21% per year as shown in Fig. 1. The spot market price of 16 Mb DRAM, for example, has dropped from US\$ 50 in January of 1995 to US\$ 2 in January of 1998, which means the average drop rate of the bit price is 60% per year. Consequently, the bit production cost of the DRAM exceeds the bit price in the time frame of the fourth quarter of 1997 to second quarter of 1998. In order to overcome the economic crisis in the semiconductor industry, chip makers can attempt to keep the bit cost below the bit price by increasing the yield (that is, improving manufacturing productivity), improving the cost effectiveness of R&D activities by modifying and upgrading the existing technology, utilizing simple processes, and developing value-added devices. More aggressively, we should focus on the development of high performance technology for higher density devices. Adopting larger wafers in production such as 300 mm, possibly in 2001, and 450 mm around the year of 2010 will be one of the major activities for the future semiconductor industry. In 1997, the bit production cost was reduced to 55% by introducing 250 nm feature size technology to replace the preceding 350 nm technology. In addition, about 20% more reduction can be achieved by introducing the 300 mm wafer technology into the production line.

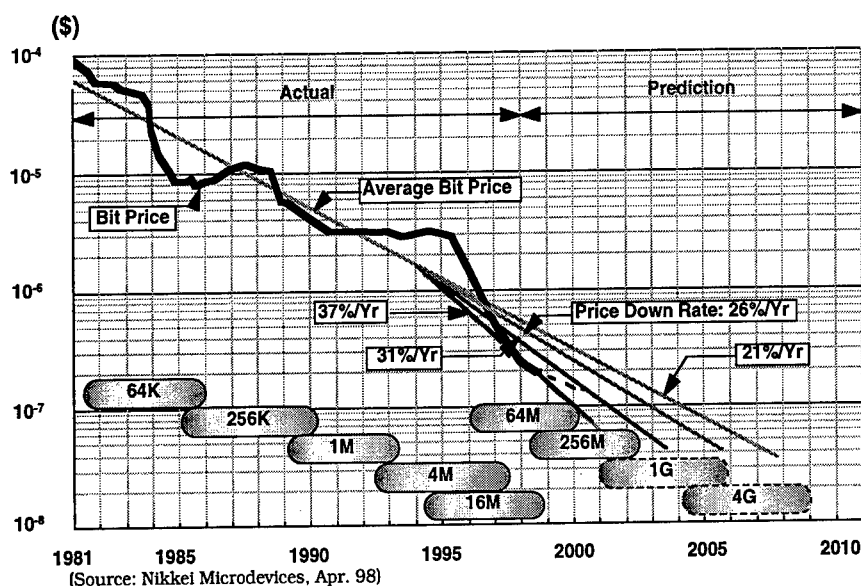


Figure 1. DRAM bit price trend.

3. Technological challenges

Even though the technological development of Si devices is facing many obstacles, it will continue until the feature size has been scaled down to 12.5 nm. The major challenges facing Si technology are optical lithography at ~ 100 nm, gate-oxide thickness at ~ 2 nm, and the increasing domination of performance by interconnects, which requires entirely new materials such as copper and low- ϵ dielectrics. In addition, the industry is also driving toward increasing the wafer diameter to 300 mm in order to reduce the bit production cost.

- *Lithography*

Lithography is one of the most rapidly advancing areas in Si technology, shifting from g- and i-line to KrF light sources, currently being used in mass production. ArF lithography is expected to support just above 100 nm feature size devices in near future. Figure 2 plots lithographic technology *versus* DRAM size and year of commercial introduction. Around the year 2010, the minimum feature size of CMOS structures will be scaled down to 50 nm, about one fifth of the current size. For feature sizes below 100 nm, the challenges in lithography can be classified into five factors: critical dimension (CD) control (resolution), overlay accuracy (alignment), economic factors including wafer throughput and cost-of-ownership, optical mask fabrication techniques, and metrology. Both optical and non-optical lithography such as X-ray, electron beam, extreme ultra-violet (EUV), ion plasma lithography (IPL) and SCALPEL are under investigation as candidates for the next generation lithography technology to enable downsizing below 100 nm. Based on a comparative evaluation of the patterning requirements for the 100 nm generation and beyond, the successful resolution should arrive in time to meet the projected production sizes of Fig. 2.

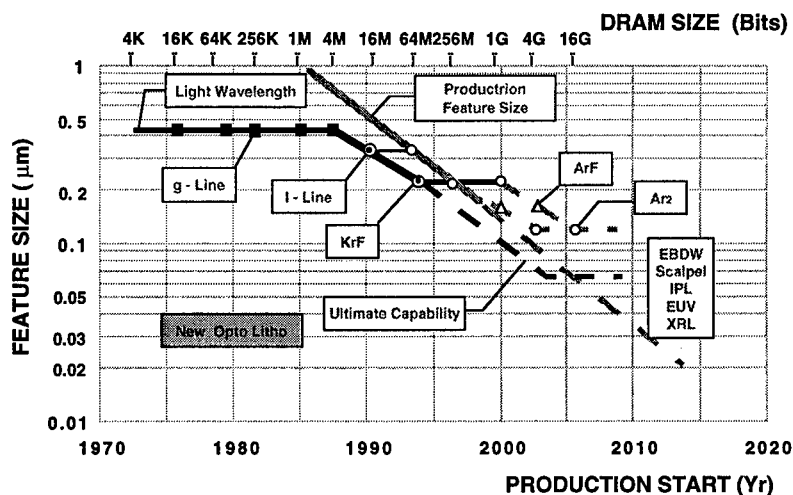


Figure 2. Lithography trend vs. manufacturing start year.

- *Front-end-of-the-line process integration*

Scaling of the MOSFET is important in meeting circuit speed goals. Current projections of speed and power requirements together with the relevant front-end-of-the-line (FEOL) technological parameters are listed in Table 1. One of the key elements of FEOL technology is the gate stack, which includes the gate dielectrics and the gate electrodes. For Si technology with a feature size of 100 nm and below, the oxide thickness will be 1.5–2 nm. Therefore, new gate dielectrics, with a higher relative dielectric constant (ϵ) than silicon dioxide ($\epsilon_{\text{ox}} \sim 3.9$) will be required. Candidate materials are silicon nitride, aluminum oxide, tantalum pentoxide, and barium strontium titanate (BST). The other key element of the gate stack is the gate electrodes. The 0.5–1 nm-thick depletion layer that forms in the polysilicon electrode increases the effective gate oxide thickness, which, in turn, reduces the gate capacitance. The use of refractory metal electrodes, such as TiN/W bi-layers, is a possible solution to avoid the formation of a depletion layer in the gate electrode. Another key FEOL element consists of the deep source/drain, including the silicide, and the ultra-shallow source/drain extension with shallow trench isolation (STI).

- *Capacitor process integration*

Memory cells need a short wire pitch of about twice the minimum feature size L (whereas a pitch of $3L$ is sufficient for logic). Projections for DRAM cell size scaling, summarized in Table 2, quickly reach the point where current dielectric material, NO films, cannot meet the required cell capacitance. Even cylindrical capacitor integration using tantalum pentoxide with TiN electrodes may not be good enough for Gb DRAMs because data retention would be extremely difficult to achieve in devices scaled below 100 nm. Therefore, a technological breakthrough involving a change in material, cell design, and memory architecture is required. Figure 3 represents the trend of stack capacitor technology in terms of materials and memory architecture. The capacitor integration of barium strontium titanate (BST) or lead zirconium titanate (PZT) with noble metal electrodes will be a solution for the giga-bit DRAM and merged memory/logic applications because of the high dielectric constant of these materials.

Year (Technology)	1998 (180 nm)	2005 (100 nm)
Clock frequency	1 GHz	2GHz
Power supply voltage	1.5–1.8 V	< 1.2 V
Chip power	80 Watt	150 Watt
I_{ON} (NMOS/PMOS)	600/280 (nA/ μm)	600/280 (nA/ μm)
Gate oxide thickness	3.5–4 nm	1.5–2 nm
Isolated line CD	150 nm	70 nm
S/D junction depth	80–150 nm	40–80 nm

Table 1. Front-end technology.

Year (Technology)	1998 (180 nm)	2005 (100 nm)
Projection area of a cell	150x300 nm	100x180 nm
Required capacitance	> 25 fF/cell	> 25 fF/cell
Dielectric materials	Ta ₂ O ₅	BST, PZT
Equivalent oxide thickness	2.5–3 nm	< 0.2 nm
Storage node height	1000 nm	< 500 nm

Table 2. Capacitor technology.

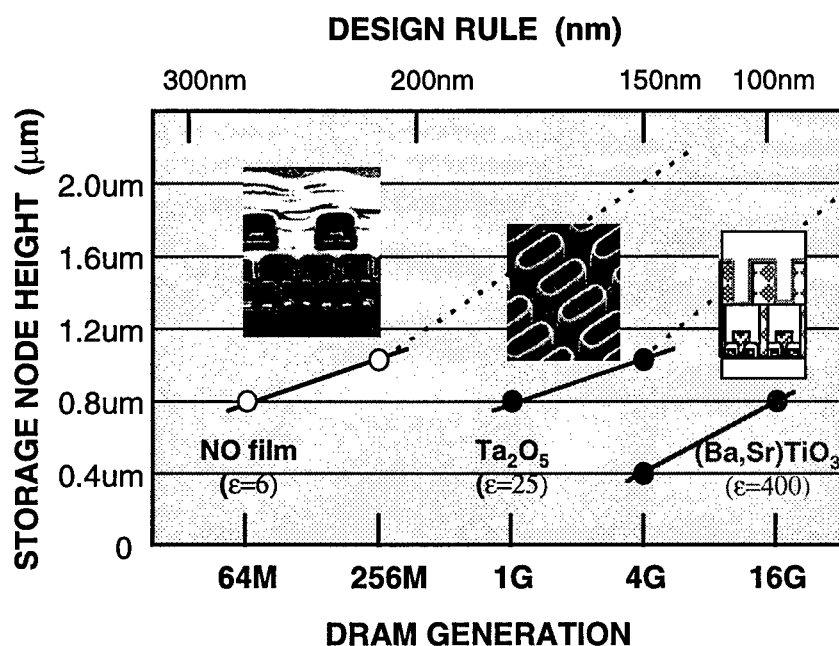


Figure 3. Trends in capacitor materials.

- *Interconnect (back-end) process integration*

To date, the standard interconnect scheme for high performance logic devices consists of four to six layers of aluminum metal lines with tungsten plugs for the vias and SiO₂ based inter-level dielectric (ILD). The projected interconnect technology requirements are shown in Table 3. With the start of production for the 150 nm technology generation, it is expected that the semiconductor industry will switch to using a lower resistivity material, Cu, for the metal lines and vias, and low-ε dielectric materials for the ILD. As shown in Fig. 4, low-ε dielectric

Year (Technology)	1998 (180 nm)	2005 (100 nm)
Number of metal levels	6	8
Minimum contact/via CD	200/250 nm	110/140 nm
Metal width	180 nm	100 nm
DRAM contact height/width aspect-ratio	6–8	10–12
Via height/width aspect-ratio	2.2	< 3
Barrier thickness	80 nm	10 nm
Interlevel dielectric ϵ	2.5–3.0	1.5–2.0

Table 3. Interconnect technology.

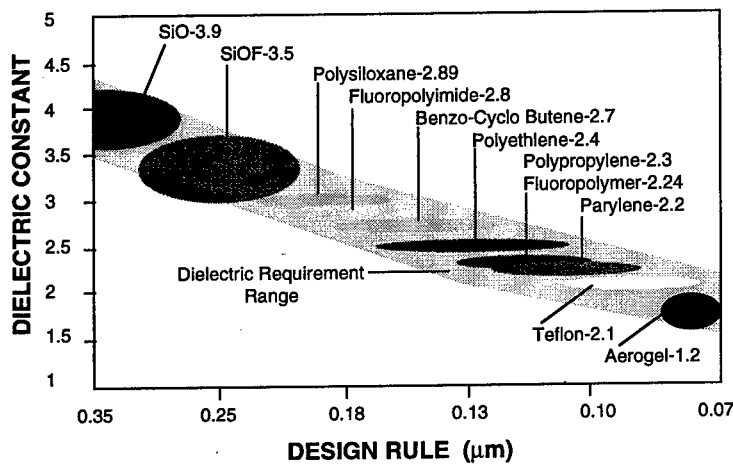


Figure 4. Trends in dielectric materials.

materials, including fluorine-doped silicon oxide, spin-on coated organic and inorganic materials, and aerogels are under active development.

It is expected that significant improvement in IC performance can be achieved by using the copper/low- ϵ ILD system because it has lower resistance and parasitic capacitance, and sustains higher current densities than the aluminum/SiO₂-based ILD system. The critical issues in device applications are electrical breakdown, thermal stability, adhesion and reliability.

- *Device architecture*

The scaling of conventional MOSFETs results in several problems arising from the reduced power supply voltage and the rapid reduction of gate-oxide thickness. The most severe problem arises from the non-scalability of threshold

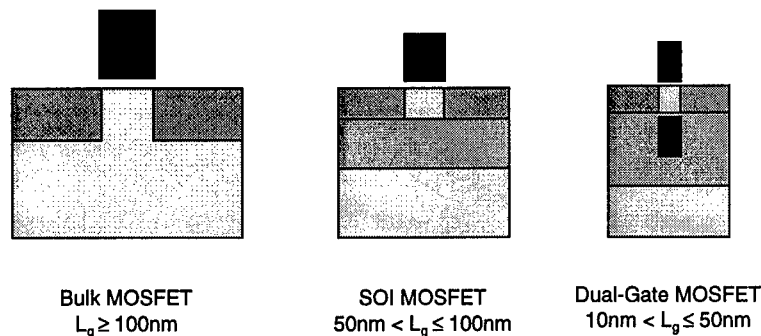


Figure 5. Evolution of the MOSFET architecture.

voltage, which leads to reduced current drive and thus degrades the circuit speed at 130 nm and below technology generations. Even though the transistor gate lengths are scaling down 30% per generation, the maximum current drive (I_{ON}) remains constant or even decreases slightly with each generation. The rapidly decreasing gate-oxide thickness creates a number of problems, such as increased off-state leakage, boron penetration, and thickness control. Therefore, the conventional oxide needs to be replaced by a new materials at the 100 nm generation. It is expected that downscaling down to the 50 nm generation will be achieved using a silicon-on-insulator (SOI) structure. The SOI structure allows the realization of ultra-shallow junctions and the significantly reduced channel doping profile due to the backgating effects of the insulator interface. Figure 5 shows the evolution of the basic MOSFET architecture from the conventional single gate bulk MOSFET, to the SOI MOSFET, and ultimately to the dual-gate MOSFET.

4. Value added devices: system-on-a-chip

The appearance of the system-on-a-chip has impact on the design synthesis. Economical integration of a number of devices on a single chip, currently growing exponentially, is of great interest, and will drive an integration between synthesis and system-level design. Typically, a system-on-a-chip is not a single chip of application-specific logic, but rather a set of naturally organized functional blocks that are combined into a special-purpose module. The enormous know-how of mainstream microelectronics is driven by memory and logic applications, and more and more by "system-on-a-chip" type applications with on-chip memory and logic functions. Moreover, multiple functions from microprocessor, ASIC, logic, memory and input/output, which can be referred to as system building blocks, or more commonly intellectual property blocks, will be integrated in a single chip. As semiconductor technology has matured, the industry has managed to implant more functions onto chips, finally creating these ambidextrous "superchips".

Superchips actually work like miniature computers, complete with processors and memories sophisticated enough to run multimedia and communication systems. The trend to on-chip memory and logic functions will cause microelectronics to go forward into systems with increasing system knowledge and value-added in silicon.

The accelerating evolution towards systems-on-a-chip and the multimedia communication market dictates that increased investment and an improved coordination between user companies, universities, and design technology suppliers are essential. This coordination will enforce much more intensive cooperation of system and semiconductor manufacturers than in the past.

5. Conclusions

Semiconductor logic and memory chips have achieved dramatic growth and development in performance and density fueled by rapid advances in technology. However, the price drop rate of semiconductor chips, especially for DRAMs, is much more precipitous than the cost drop rate. It dictates that the industry should concentrate its resources on high-performance technology for higher density devices, lower cost R&D activities, and high productivity. The driving forces for the semiconductor industry of the future will include, but not be limited to, minimization of production cost with larger wafers, technology breakthroughs that require non-prohibitive development cost, and value-added devices for higher performance.

Future Trends in Large-Scale Integrated Circuit Technologies from an Industrial Perspective

Yoichi Unno and Hiroshi Iwai

Microelectronics Engineering Laboratory, Toshiba Corporation, Kawasaki, Japan

1. Introduction

Has the downsizing of large-scale integrated (LSI) circuits reached its limits? From the device standpoint, the operation of small sub-0.1 μm CMOS circuits has already been experimentally confirmed.¹ However, with regard to large LSI circuitry, an important question arises: can circuits with such ultra-small dimensions be fabricated by the extension of current device/process production technologies? The biggest issues are whether the lithography and interconnects technologies can realize such an ultra-small dimensions in production.

Lithography is facing the challenge of optical resolution limitations. Fortunately, it is already recognized that 0.15 μm critical dimension (CD) can be achieved by the KrF laser, and furthermore, it is believed that ArF laser with an even shorter wave length can realize sub-0.13 μm resolution. Unfortunately, however, it is not clear at this moment what is the most likely technology for lithography at even smaller dimensions. Although great efforts have been put into research and development related to lithography based on X-ray, electron-beam, and EUV technology, the prospects for economic efficiency and reliability are still unknown for all of these candidates.

In the meantime, difficulties with interconnect technology have increased as the miniaturization of LSI has progressed. Basically, CMOS is a technology with low power consumption; nevertheless, the current density in interconnects has already reached 10^5 A/cm^2 . This is already close to the limitation of today's aluminum wiring and further improvements to the current capacity have been attempted. For this reason, new materials for interconnects, such as copper, have been aggressively pursued, because aluminum may not satisfy the current capacity requirements of near future high-end CMOS MPUs.

The question from an industrial viewpoint is whether LSI circuits fabricated using such sophisticated technologies can be delivered at an acceptable cost. As is obvious from the case of DRAM manufacturing, there is very severe price competition in the semiconductor marketplace, making it difficult to recoup the initial capital investment. In fact, plant and equipment investment on the conventional model are already facing great difficulties.

This article examines the expected difficulties of the future LSI industry in terms of technology and economy. Possible future directions for LSI development are also discussed.

2. Trend in LSI growth

Large-scale integration has made remarkable progress over the past 30 years. The first generation of LSI circuits, which appeared as products such as a 1 Kbit DRAM and a 750 kHz microprocessor unit (MPU), has evolved up to 64 Mbit DRAMs and 600 MHz MPUs, as shown in Table 1. This evolution has seen the critical dimension (CD) decrease by a factor of 40, the number of memory bits increase by 64,000 times, and the MPU clock frequency increase by a factor of 800. Recent revolutionary progress in the information and communication fields, as represented by the Internet, and the popularization of intelligent mobile devices owe much to this remarkable development of LSI technologies.

Year	CD	ratio	DRAM bit	ratio	MPU clock	ratio
1970-72	10 μm	1	1K (Intel 1103)	1	750 kHz (Intel 4004)	1
1998	0.25 μm	1/40	64M (many vendors)	64,000	600 MHz (DEC Alpha 21264)	800

Table 1. Evolution of LSI from 1970 to 1998 (CD stands for critical dimension).

Year	1970	1974	1977	1980	1983	1986	1989	1989	1995	1998
CD (μm)	10.0	6.0	4.0	3.0	2.0	1.2	0.8	0.5	0.35	0.25
Shrink Rate	--	0.6	0.67	0.75	0.67	0.6	0.67	0.63	0.70	0.71
DRAM	1K	4K	16K	64K	256K	1M	4M	16M	64M	256M

Table 2. Historical trends in critical dimension (CD) and production DRAM size.

Year	1997	1999	2001	2003	2006	2009	2012
Dense Lines: Half pitch (μm)	0.25	0.18	0.15	0.13	0.10	0.07	0.05
Shrink Rate	—	0.72	0.83	0.86	0.77	0.7	0.71
Isolated Lines: MPU gate (μm)	0.20	0.14	0.12	0.10	0.07	0.05	0.035
Shrink rate	—	0.70	0.86	0.83	0.70	0.71	0.70
DRAM @samples/introduction	256M	1G	—	4G	16G	64G	256G
DRAM @production ramp	64M	256M	1G	1G	4G	16G	64G
MPU Clock Frequency (MHz)	750	1200	1400	1600	2000	2500	3000

Table 3. Future trends forecast by the 1997 SIA roadmap.

The tremendous growth of LSI circuitry shown in Table 2 has been achieved by the downsizing of components based on the Moore's Law,² in which CD reduces to 2/3, chip size increases to 3/2, and number of components in a chip increases by 4 times every 3 years or every new generation. Regarding the future, a similar trend as the historical one was expected by the Semiconductor Industry Association 1997 Roadmap³ as shown in Table 3 and continuous growth of LSI is forecast to reach the 0.05 μm generation in the year 2012.

If we extrapolate the same trends to the year 2060, the CD of LSI products would bottom out near the atomic lattice constant of silicon, leading to the production of the 64 Ebit (10^{18}) DRAM — see Table 4. This would be the ultimate limit of LSI. Before reaching this limit, however, we can expect certain other limitations to arise the reasons shown in the table. In particular, the 0.1 μm generation is thought to be a critical stage because four major downsizing limitations coincide: performance, lithography, interconnects, and cost. In this article, we will examine these factors individually and then comment on their interrelated nature.

3. Performance limits to downsizing

Progress in LSI circuit performance has been achieved by the downsizing of components, mainly because the capacitance values decrease, leading to faster switching times. This may no longer be true when the CD decreases to 0.1 μm and below. First of all, the interconnect capacitance between the 2 parallel metal lines increases abruptly as the separation falls below a few times 0.1 μm . The problems related with the interconnects are explained in the interconnects section. Furthermore, performance improvements in MOSFET devices cannot be guaranteed in the generations of 0.1 μm and below.

Table 5 shows the current status of the gate lengths for each level of front-end research and products. It should be noted that the minimum transistor L_G achieved thus far ($L_G = 0.04 \mu\text{m}$) does not give the highest transistor performance (which

Year and Product	2006	2018	2027	2048	2060
CD	0.1 μm (= 100 nm)	0.025 μm (= 25 nm)	0.01 μm (= 10 nm)	0.001 μm (= 1 nm)	0.00025 μm (= 2.5 Å)
DRAM	16G bit (Giga: 10^9)	256G bit	16T bit (Tera: 10^{12})	256P bit (Peta: 10^{15})	64E bit (Exa: 10^{18})
Downsizing Limiting Factors	Performance? Lithography? Interconnects? Economy?	MOSFET switch-off?	Thermal noise? Uncertainty principle?		Atomic distance

Table 4. Simple extrapolation of historical trends into the future and expected downsizing limiting factors.

Device Level	Min. L_G (Simulation)	n -MOS: $0.025 \mu\text{m}^4$	p -MOS: $0.025 \mu\text{m}^5$
	Min. L_G (Experiment)	n -MOS: $0.04 \mu\text{m}^6$	p -MOS: $0.05 \mu\text{m}^7$
	L_G for max I_D (Exp.)	n -MOS: $0.06 \mu\text{m}$, $1.8\text{mA}/\mu\text{m}$ @ 1.5V^8	
	L_G for max g_m (Exp.)	n -MOS: $0.06 \mu\text{m}$, $1120\text{mS}/\text{mm}$ @ 1.5V^8	
		n -MOS: $0.07 \mu\text{m}$, $>1100\text{mS}/\text{mm}$ @ 1.5V^9	
		n -MOS: $0.10 \mu\text{m}$, $1210\text{mS}/\text{mm}$ @ 2.5V^{10}	
Circuit Level	Min. L_G (Experiment)	CMOS Ring Osc.: $0.075 \mu\text{m}$, 22ps @ 1.5V , 16ps @ 3.0V^1	
	L_G for min tpd (Exp.)	CMOS Ring Osc.: $0.1 \mu\text{m}$; 8.0ps @ 2.5V^{11}	
		SOI CMOS Ring Osc.: $\sim 0.1 \mu\text{m}?$ ($L_{\text{eff}} = 0.07\mu\text{m}$), 7.9ps @ 2.1V^{12}	
LSI Level	Min L_G for MPU (Production)	$0.25 \mu\text{m}$: 400MHz @ 2V Pentium II ¹³	
	L_G for max clock MPU (Production)	$0.35 \mu\text{m}$: 600MHz @ 2V Alpha 21264 ¹⁴	

Table 5. Gate lengths which give the highest performance for digital logic.

occurs at $L_G = 0.06\text{--}0.07 \mu\text{m}$ instead), and that the minimum L_G achieved for in a CMOS ring oscillator ($L_G = 0.075 \mu\text{m}$) does not give the best circuit performance (which is obtained at $L_G = 0.1 \mu\text{m}$). Clearly, thus far the minimum L_G for MOSFET operation has not guaranteed the best transistor or circuit-level performance because of factors like the high resistance due to ultra-shallow S/D junctions. With logic devices, the increase in power consumption with rising clock frequency is another major concern. Reducing the supply voltage is the most effective way to reduce the power, though at the cost of reduced clock speed.

It should also be noted that no circuit performance improvement has been realized below the $0.1 \mu\text{m}$ generation, even in the laboratory, despite the availability of deep sub- $0.1 \mu\text{m}$ gate length MOSFETs. This would appear to present a downsizing limit in terms of LSI performance. We hope this will be solved by future improvement of downsizing technology or by some breakthrough.

4. Lithographic limits to downsizing

Lithography is the fundamental process by which downsizing takes place. To date, progress in lithographic resolution has been achieved by reducing the wavelength of the light used to expose the resist. Currently, $0.248 \mu\text{m}$ (248nm) wavelength light, as emitted by a KrF excimer laser, is being used for $0.25 \mu\text{m}$ devices, as shown in Fig. 1. It should be noted that KrF will also be used even for

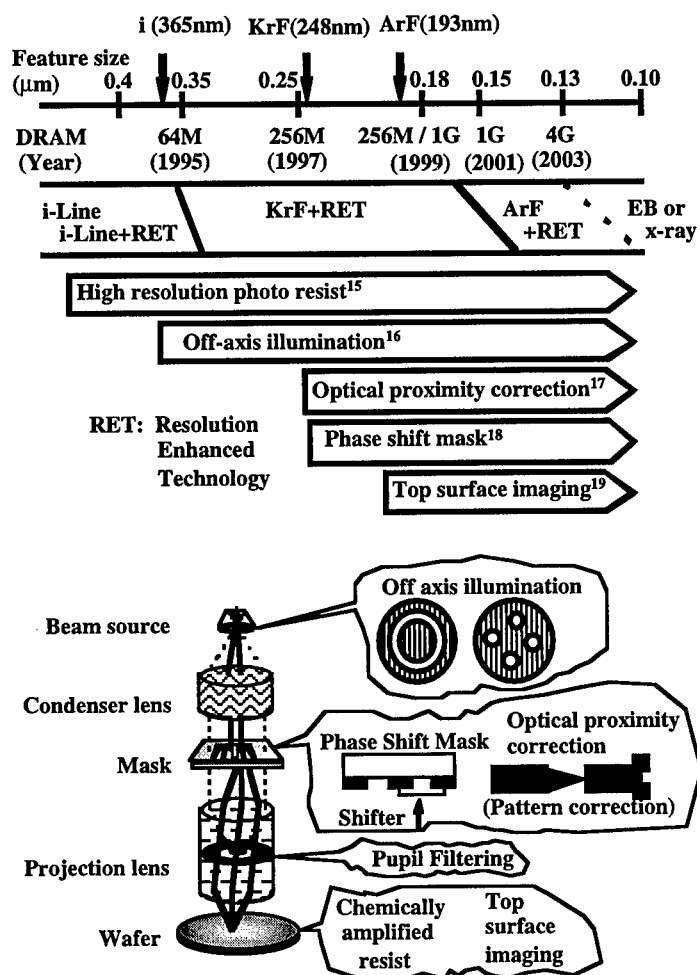


Figure 1. Resolution enhancement techniques.

0.15 μm lithography, which is much shorter than the wavelength. In order to realize this, several types of resolution enhancement techniques (RET) have been developed or under development, as also shown in Fig. 1. Furthermore, it is expected that 0.193 μm (193 nm) wavelength ArF light combined with RET will push lithography down to 0.13 μm and probably down to 0.1 μm in the near future.

So far it appears that 0.13 μm (and probably 0.1 μm) generation can be handled by the ArF stepper. One of the most serious problems, however, is that there is no clear candidate commonly accepted by the industry to replace ArF sources below 0.1 μm. Several lithographic techniques have been proposed, as shown in Table 6, but none has yet been confirmed as a cost-effective technology. We hope this quandary will be solved in the near future.

VUV (Very Ultraviolet) ²⁰	EUV(Extreme Ultraviolet) ²¹	X-ray Proximity ²²	Electron Beam: Cell Projection ²³ or SCALPEL ²⁴	Ion Projection Lithography ²⁵
--------------------------------------	--	-------------------------------	---	--

Table 6. Candidates for lithography for 0.1 μm and below (SCALPEL stands for scattering with angular limitation projection e-beam lithography).

5. Interconnect limits to downsizing

Interconnects are another major concern because the length of the interconnects, and hence the number of interconnect layers, continues to increase with every LSI generation, as illustrated in Fig. 2. Thus, the number of back-end process steps increases, resulting in a significant rise in production costs. Because the interconnect lines become longer, narrower, and denser with every generation, their resistance and capacitance increase dramatically. The resulting RC delays and the production cost of such interconnects will certainly act as limiting factors on LSI downsizing in the regime of 0.1 μm and below.

Of course, the introduction of new techniques such as Cu-damascene and low- κ interlayer materials would effectively extend the limit (see Fig. 3). A schematic 0.1 μm interconnect structure is shown in Fig. 4, but its complexity convinces us that the limitation imposed by interconnect multi-layering and downsizing will be reached in the not too distant future.

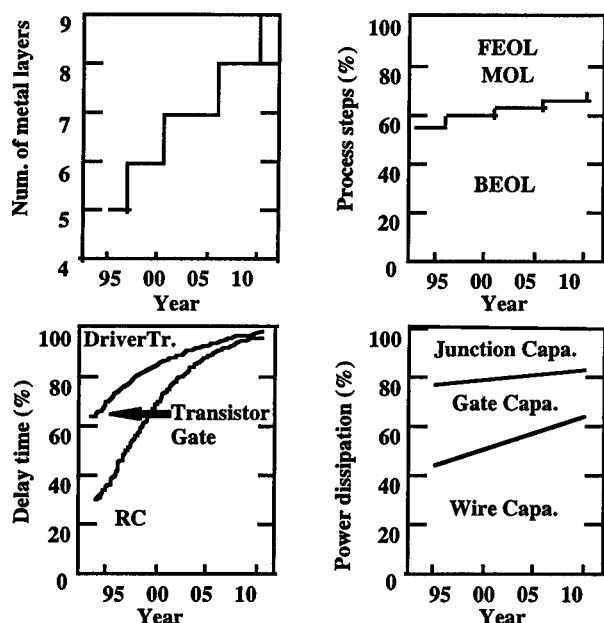


Figure 2. Several aspects of interconnect trends.²⁶

Feature size (μm)	1995 0.50	1997 0.35	1999 0.25	2003 0.18/0.15	2003 0.13
# of Interconnects	3	3~5	4~6	6~7	6~7
Interconnect	Al-Cu		Cu		
Plug	W	Al-Damascene ²⁷		Cu-Damascene ^{28,29}	
ILD	TEOS $\kappa = 4.1$	F-TEOS \rightarrow FSG $\kappa = 3.7\sim 3.3$ ³⁰		Organic $\kappa \leq 3$ ³¹	
Barrier Metal	Collimated Ti/TiN	CVD TiN \rightarrow CVD Ti/TiN			
Planarization	LT reflow	CMP ³²			

Figure 3. Trends in interconnect technology.²⁶

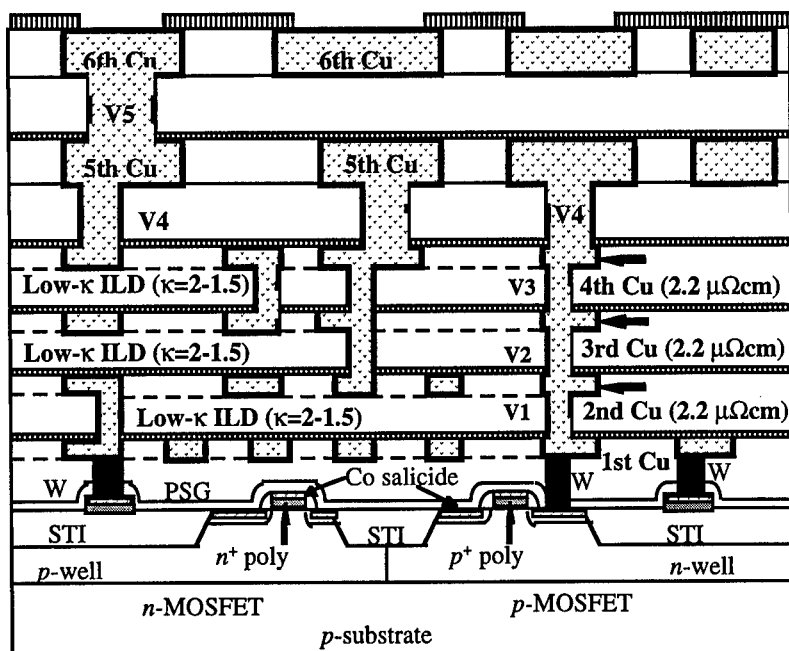


Figure 4. 0.1 μm Cu and low- κ interconnection structure.²⁸

6. Cost limits to downsizing

As noted in the previous sections, the number of process steps increases with each new generation, and each step requires sophisticated and accurate control. This leads to significant production cost increases. It should be noted that the fraction of total LSI cost due to production equipment has kept rising over recent years. Figure 5 compares semiconductor equipment sales to the LSI circuit sales. It shows that the equipment costs, estimated from the equipment to LSI sales ratio, have risen to around 20% of the total LSI market from only 10% a decade ago. The equipment cost fraction is expected to rise to 30% in the year 2005 and further to 50% in the year 2020. Moreover, the investment involved in building a single advanced LSI fabrication line will reach several billion dollars in 20 years if a simple extrapolation based on a 25% increase at every generation, as shown in Table 7. Thus, the cost issue will certainly inhibit LSI progress in the near future and might even suppress LSI growth for a certain period.

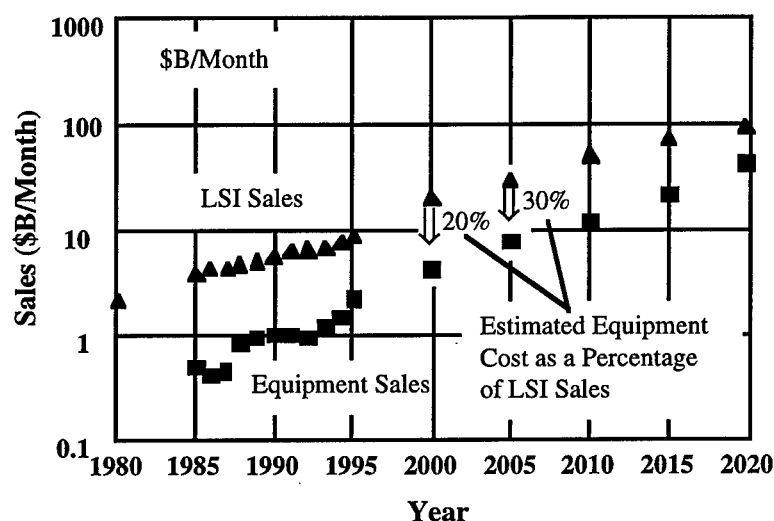


Figure 5. Sales trend of LSI and LSI equipment.³³

Year	97	99	01	03	06	09	12	15	18
CD (μm)	0.25	0.18	0.15	0.13	0.10	0.07	0.05	0.035	0.025
Investment (B\$)	1	1.25	1.56	1.95	2.44	3.05	3.81	4.78	6.00

Table 7. Simple extrapolation of the investment trends for LSI fabrication, assuming a 1 B\$ cost of a 64 M-DRAM fab and an increase rate of 25% for every generation.

7. Conclusion and prospects for the future

Four possible limiting factors on LSI progress down to the 0.1 and sub-0.1 μm generations have been described. The first three were technology-oriented issues of performance, lithography, and interconnects. In the 0.1 and sub-0.1 μm generations, however, each technological issue ultimately comes back to a cost question, given the fact that at least some (primitive) research-level techniques for realizing such devices have been already reported. In that sense, cost is the most fundamental of these four factors.

What will be the future direction of technology development in general? Will it leave the beaten path? Will there be any major breakthrough? No one can make predictions about future breakthroughs, but it seems that an immediate revolution on the hardware side is unlikely. Rather than hardware, there is a better chance of a breakthrough, or at least of innovation, on the software side, such as new applications, algorithms and system architectures. The Internet mentioned in Section 2 would be one of the examples. It is one of the major forces to expedite the realization of the information community and thus to develop new application fields for LSI circuitry. Popularization of the intelligent mobile devices is another example.

In particular, it has been noted that today's computer architectures are much less efficient than biological structures. Thus, there is room for revolutionary improvements in the efficiency of our computer structures at some point in the future. Once a breakthrough or innovation occurs on the software side, hardware will follow. New hardware demands imposed by new systems will become a strong motive force. In fact, the past development of digital circuits has strongly depended on the conventional computer architectures. The discussion of the necessity of the downsizing in the SIA Roadmap has been made also on the premise of today's architectures. Yet it can be expected that new algorithms and architectures, perhaps totally different from those in use today, will promote the development of a completely new style of LSI circuits. Even now, prototypes of future hardware device components might already be in existence simply waiting for suitable applications.

Research on conventional path is still crucial. It is likely that research on the introduction of many kinds of new materials — whether inorganic, organic, or even biological — to semiconductor devices will lead to productive innovations. This certainly is an exciting area with potential for circumventing some limitations of LSI circuitry. From the process research point of view, we look forward to new approaches that will dramatically reduce the production costs. Also, environmental issues must be considered of utmost importance in the development of any new techniques.

Traditionally, the LSI industry has chosen to pursue lower chip costs by betting heavily on the downsizing of devices. This, however, has often caused an over-supply of DRAMs and led to rapid collapse of market prices in spite of the huge investment for realizing LSI downsizing. This paradigm has often harmed the sound growth of the LSI industry.

For the development of the 21st century electronics industry, sound growth of the LSI and LSI-related industries is critical. In order to avoid useless competition and thus to make an efficient development of technologies, cooperation on the global level will be very important. During the 1990s, alliances between competing LSI manufacturers in Japan, the US, and Europe and further in Korea, Taiwan, and Singapore became a popular way to share the huge cost of R & D. On the other hand, a silicon foundry business specializing in the production of LSI circuitry has arisen in Taiwan during 1990s. The trend in the 21st century will be for conventional comprehensive LSI manufacturers to evolve into (or be replaced by) technically specialized companies because of the sophistication in manufacturing technology accompanied by extreme downsizing, the difficulty of highly-specialized circuit design techniques, and the creativity requirements of building new systems. However, because of the trade-off between technological specialization and huge development costs, alliances between companies will become much more complex as mixed vertical and horizontal structures develop.

Regardless of the shape of the LSI industry in the 21st century, advances in technology are the key to LSI progress. Since it will prove impossible for the huge corporate-level laboratories owned by companies to afford such development by themselves, cooperation between the industry and academia on a global scale is inevitable, and this will demand a deep mutual understanding among those involved in such collaboration.

The ultimate purpose of all technological development is to contribute to human society by enriching our life. The development of the LSI technology is no exception. In particular, in the near future, industrial societies with aging demographic profiles and insufficient working populations will undoubtedly depend on LSI circuits to support and even replace human workers in an ever expanding variety of occupations, contributing to the overall welfare of society. Thus we are looking forward to the evolution of LSI circuits to safeguard a great future for humanity.

8. Acknowledgments

The authors would like to thank Drs. Hideki Shibata and Ichiro Mori of Toshiba's Microelectronics Engineering Laboratory for offering figures and references regarding advanced interconnect and lithography technologies.

References

1. T. Yamazaki, K. Goto, T. Fukano, *et al.*, "21 psec switching 0.1 μm -CMOS at room temperature using high performance Co salicide process," *IEDM Tech. Digest* (1993), p.906.
2. G. E. Moore, "Progress in digital integrated circuits," *IEDM Tech. Digest* (1975), p. 11.

3. Semiconductor Industry Association, *The National Technology Roadmap for Semiconductors*, 1997.
4. C. Fiegna, H. Iwai, T. Wada, *et al.*, "A new scaling methodology for the 0.1–0.025 μm MOSFET," *Symp. VLSI Technol.* (1993), p. 33.
5. C. Fiegna, H. Iwai, T. Wada, *et al.*, "Scaling the MOS transistor below 0.1 μm : methodology, device structures, and technology requirements," *IEEE Trans. Electron Dev.* **41**, 941 (1994).
6. M. Ono, M. Saito, T. Yoshitomi, *et al.*, "Sub-50 nm gate length *n*-MOSFETs with 10 nm phosphorus source and drain junctions," *IEDM Tech. Digest* (1993), p. 119.
7. A. Hori, H. Nakaoka, H. Uemimoto, *et al.*, "A 0.05 μm -CMOS with ultra shallow source/drain junctions fabricated by 5keV ion implantation and rapid thermal annealing," *IEDM Tech. Digest* (1994), p. 485.
8. G. Timp, A. Agarwal, F. H. Baumann, *et al.*, "Low leakage, ultra-thin gate oxides for extremely high performance sub-100 nm *n*-MOSFETs," *IEDM Tech. Digest* (1997), p. 930.
9. A. Chatterjee, R. A. Chapman, G. Dixit, *et al.*, "Sub-100 nm gate length metal gate NMOS transistors fabricated by a replacement gate process," *IEDM Tech. Digest* (1997), p. 821.
10. H. S. Momose, S. Nakamura, T. Ohguro, *et al.*, "A study of hot-carrier degradation in *n*- and *p*-MOSFETs with ultra-thin gate oxides in the direct-tunneling regime," *IEDM Tech. Digest* (1997), p. 453.
11. D. Hisamoto, K. Umeda, K. Ohnishi, J. Yugami, and T. Shiba, "Sub-10-ps gate delay by reducing effect ant an extension" *IEDM Tech. Digest* (1997), p. 239.
12. F. Assaderaghi, W. Rausch, A. Ajimera, *et al.*, "A 7.9/5.5 psec room/low temperature SOI CMOS," *IEDM Tech. Digest* (1997), p. 415.
13. Technical specifications at <http://www.intel.com>.
14. Technical specifications at <http://www.digital.com>.
15. H. Ito and C. G. Willson, "Applications of photoinitiators to the design of resists for semiconductor manufacturing," in: T. Davison, ed., *ACS Symposium Series No. 242*, Washington, D.C.: ACS, 1984, p.11.
16. N. Shiraishi, S. Hirukawa, Y. Takeuchi, and N. Magome, "New imaging technique for 64M-DRAM," *Proc. SPIE* **1674**, 741 (1992).
17. A. Starikov, "Use of a single size square serif for variable print bias compensation in microlithography : method, design, and practice," *Proc. SPIE* **1088**, 34 (1989).
18. M. D. Levenson, N. S. Visnawathan, and R. A. Simpson, "Improving resolution in photolithography with a phase-shifting mask," *IEEE Trans. Electron Dev.* **29**, 1828 (1982).
19. G. N. Taylor, L. E. Stillwagon, and T. Venkatesan, "Gas-phase-functionalized plasma-developed resists: initial concepts and results for electron-beam exposure," *J. Electrochem. Soc.* **131**, 1658 (1984).

20. T. M. Bloomstein, M. W. Horn, M. Rothschild, *et al.*, "Lithography with 157 nm lasers," *J. Vac. Sci. Technol. B* **15**, 2112 (1997).
21. H. Kinoshita, K. Krihara, Y. Ishii, and Y. Torii, "Soft x-ray reduction lithography using multilayer mirrors," *J. Vac. Sci. Technol. B* **7**, 1648 (1989).
22. D. L. Spears and H. I. Smith, "High precision pattern replication using soft x-rays," *Electronics Lett.* **8**, 102 (1972).
23. Y. Sakitani, H. Yoda, H. Todokoro, *et al.*, "Electron-beam cell-projection lithography," *J. Vac. Sci. Technol. B* **10**, 2759, (1992).
24. S. D. Berger and J. M. Gibson, "New approach to projection-electron lithography with demonstrated 0.1 μm linewidth," *Appl. Phys. Lett.* **57**, 153 (1990).
25. G. Gross, "Ion projection lithography: next generation technology?" *J. Vac. Sci. Technol. B* **15**, 2136 (1997).
26. Hideki Shibata, private communication.
27. C. W. Kaanta, S. G. Bombardier, W. J. Cote, *et al.*, "Dual damascene: a ULSI wiring technology," *Proc. VLSI Multi-Level Interconnect Conf.*, 1991, p.144.
28. S. Venkatesan, A. V. Gelatos, V. Mirsa, *et al.*, "A high performance 1.8V, 0.2 μm CMOS technology with copper metallization," *IEDM Tech. Digest* (1998), p.769.
29. D. Edelstein, J. Heidenreich, R. Goldblatt, *et al.*, "Full copper wiring in a sub-0.25 μm CMOS ULSI technology," *IEDM Tech. Digest* (1998), p.773.
30. H. Miyajima, R. Katsumata, N. Hayasaka, and H. Okano, "Formation mechanism of F-added SiO_2 films using plasma CVD," in: *16th Symp. Dry Processing*, Tokyo, Japan, 1994, p.133.
31. S. W. Russell, A. J. McKerrow, W.-Y. Shin, *et al.*, "Integration challenges of ultra low- κ dielectrics," in: *MRS Proc. Advanced Metall. Interconnect Systems ULSI Applications* (1997), p. 289.
32. P. A. Buke, "Semi-empirical modeling of SiO_2 chemical-mechanical polishing planarization," *Proc. IEEE VMIC* (1991), p. 379.
33. Private communication from Semiconductor Leading Edge Technology, Inc. (SELETE).

Driving Factors and Breakthroughs for Higher Performance Semiconductor Devices

H. Watanabe

R&D Group, NEC Corporation, Kawasaki 214-8555 Japan

1. Semiconductor market growth

Since 1970 the semiconductor industry has constantly grown at an average rate of 15–17% per year, reaching about \$150 billion worldwide in 1997, with compound semiconductor devices holding about 3.5% of the total. By exploiting new fields of application and novel added-value features with an increasing number of integrated transistors, silicon large-scale integration (LSI) performance has increased drastically in tandem with chip size growth. Demand for volume production of ever-larger chips has been reflected in a generation-by-generation increase of Si wafer diameter. As shown in Fig. 1, the total sales of Si devices exceed those of Si wafers by a factor of 35, while the corresponding ratio for III-V devices is only 10. The difference in the two ratios between device and wafer markets reflects the fabrication cost difference.

The various sectors of the semiconductor marketplace can be grouped by average device cost. The fabrication of III-V devices is much simpler than that of Si LSIs. However, as shown in Fig. 2, prices per unit area of III-V devices are higher than those of Si. The reliability of III-V devices used in the trunk lines of

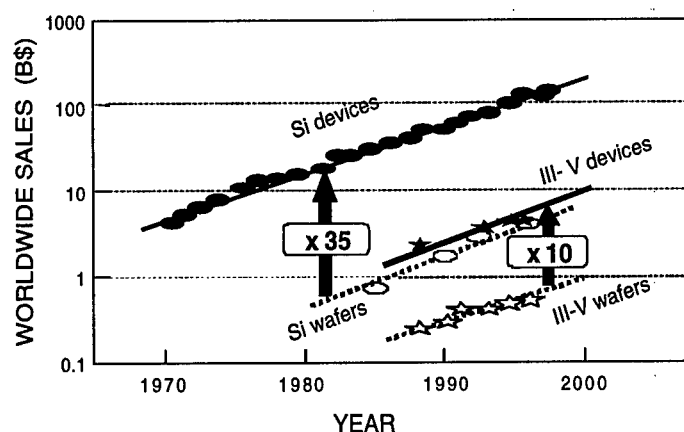


Figure 1. Worldwide sales of Si LSI, III-V devices, Si wafers, and III-V wafers.

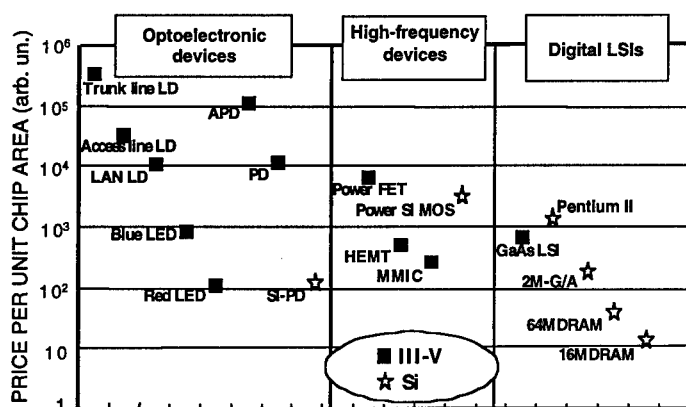


Figure 2. Price comparison of Si LSI and III-V devices.

communication systems must be as high as possible. Light-emitting diodes (LEDs) are 3–4 orders of magnitude cheaper than the trunk line laser diodes (LD). High frequency microwave devices are positioned between optical communication devices and digital LSIs. Finally, in digital silicon LSIs, logic chips are generally more expensive than memory chips, with DRAM occupying the lowest rung because of the huge market demand and an oversupply of production companies.

2. Driving factors for semiconductor device development

• Technology drivers in Si LSI development

The combination of DRAM and MPU, which are key processing components in computers, was born in 1970–1971. Since then, increased DRAM capacity and MPU clock frequency have been major demands of computer designers, as summarized in Fig. 3. The largest capacity DRAM plays the role of a technology driver, employing the most advanced feature-size miniaturization. After every new generation of DRAM, other kinds of LSIs are cost-effectively developed by modifying the DRAM-driven technology. DRAM is a density-oriented LSI, requiring highly advanced microfabrication technology, such as lithography and dry etching processes. In order to obtain as many DRAM chips as possible from a single wafer, the diameter of Si wafers has been steadily increasing.

In about 1985, personal computer (PC) systems made the jump from 16 bit to 32 bit processing and GUIs (graphic user interfaces) were introduced. Since then, the MPU has been recognized as another technology driver that has a speed-oriented culture. For high speed logic LSIs, not only shrinking the transistor size but also increasing the clock frequency has become a major goal. Various technologies for the reduction of propagation delays due to parasitic capacitance and resistance (RC delays) of interconnects have been developed with multilevel

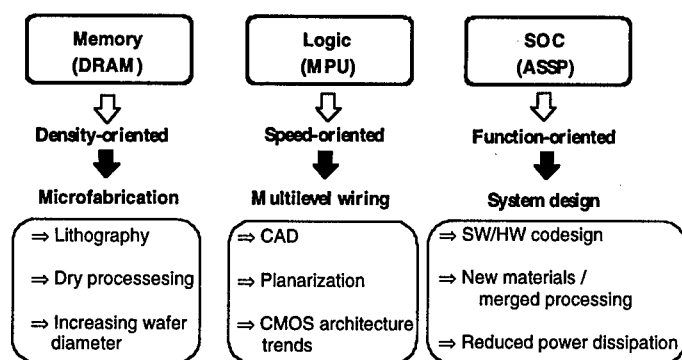


Figure 3. Technology drivers in Si LSI development.

wiring structures based on planarization processes like chemical-mechanical polishing (CMP). Further, advanced computer-aided design (CAD) tools have been developed for CMOS architecture design in logic LSIs.

System-on-a-chip (SOC) is expected to become the next technology driver with a function-oriented culture. Although integration of logic and memory has been an obvious goal for quite a long time, it drives up the costs due to an enlarged chip size and longer required design time. However, improvements in processing speed or power dissipation have become difficult to achieve in separate chips. The suppression of these problems by simple integration of memory and logic is one of the current activities of LSI engineers. Introduction of new processes using new materials has been attempted to reduce the complexity of the fabrication process.

Differences in MOS transistor processing between logic and memory and the difference in the thickness of multilevel structures in logic and DRAM should be eliminated. In order to reduce the SOC design time, simultaneous development of software and hardware (SW/HW co-design) is highly desirable for optimized integration of logic and memory blocks. We need a highly intelligent design tool at the most upstream level of LSI design — system synthesis.

- *Bandgap engineering in III-V device development*

The basic technology of compound semiconductors is quite different from Si technology. III-V compound semiconductor materials have unique properties superior to Si, such as higher mobility and direct bandgaps. High affinity between different III-V materials is one of the biggest advantages, as it enables us to design various heterostructure devices. This advantage may be the reason why integration engineering for III-V devices was delayed, with bandgap engineering occupying the central effort of III-V device engineers. Carrier confinement by double heterostructures, artificial control of electronic characteristics of electrons (holes) by superlattices, and spatial separation of carriers and donors (acceptors) by modulation-doped structures have opened a number of new applications.

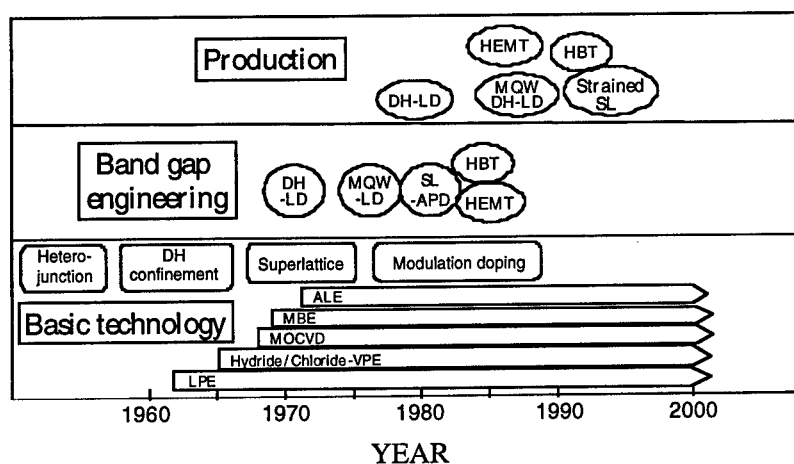


Figure 4. Driving forces in III-V device development.

Continuous improvement in epitaxial growth capability has contributed in realizing those new device ideas. Thus, strained-layer superlattices are finding application in optoelectronic and microwave devices, while three-dimensional carrier confinement by quantum dots is intensively studied for the next generation optoelectronic devices. Single-electron devices and their physics and fabrication are becoming very popular. Uniquely, compound semiconductor research spans the entire range from quantum solid-state physics to commercial devices. A summary of the driving forces in III-V research is shown in Fig. 4.

3. Technology breakthroughs

In the early stages of device development history, the main effort was devoted to controlling the electronic properties of a single junction: *pn* junction, MOS oxide-semiconductor interface, Schottky contact, or a single heterojunction. In the case of Si, research evolved towards integration engineering, while III-V research came to focus mainly on bandgap engineering. As a result, the techniques aimed at increasing added value in a device are completely different. Silicon is intensively moving towards ultra-high-density integration (ULSI or SOC) in a single chip, where the key issue is efficient performance design. One method of reducing costs is the joint utilization of a circuit block (often known as an intellectual property block) developed via worldwide collaboration of semiconductor companies. Multi-chip modules (MCMs) and three-dimensional packaging are also proposed to suppress the cost increase. For the next generations of ULSI, several challenges such as on-chip parallel, multivalued, or neural computing are under study as new information processing ULSI architectural paradigms.

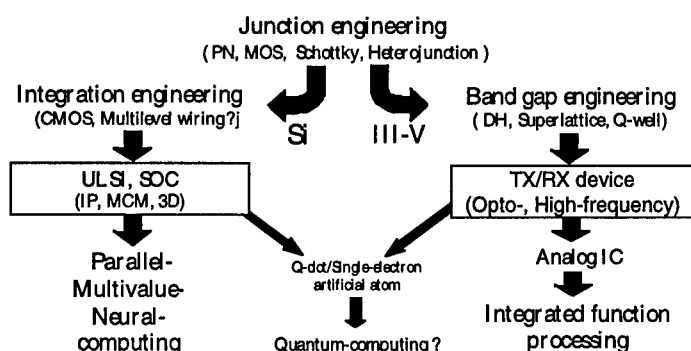


Figure 5. Technology breakthroughs in Si and III-V devices.

On the other hand, with the exception of a small number of ASICs, III-V devices in the marketplace are discrete transmitter or receiver devices for optoelectronic and microwave applications. Small-scale analog integration has been started for cellular phone electronics. Several approaches to increase functionality by combining different types of devices are currently under development. Integration of optoelectronic semiconductor devices with light waveguides using dielectric films is a current effort to reduce the module cost.

Quantum dot device physics and its applications are being studied in both the Si and III-V fields. The gate lengths in both Si MOSFETs and GaAs MESFETs are already as low as about 0.18–0.10 μm in commercially available chips, while research prototypes have been fabricated down to approximately 0.01 μm . The resulting quantum dot structures are studied by both Si and III-V researchers as potential single electron devices — a convergence illustrated in Fig. 5. Wavefunction engineering, which includes artificial shaping of wavefunctions, artificial interconnection between wavefunctions and orbital modulation of interconnected wavefunctions, may be promising for basic system architecture — for example, using quantum-computing methods. The device applications of quantum dots as ultimate ultra-low-power devices are still in their infancy.

4. Future target devices

For multimedia Internet communication the processing speed of the system must be further improved. Future device performance is typically discussed in the context of roadmap planning. The roadmap forecasts device improvement by projecting into the future the step-by-step evolution of presently used concepts. For instance, the processing performance of MPUs will be improved by increasing the number of transistors. However, this increase entails larger chip sizes and power dissipation, as well as increased design time. In turn, these increases require more resources: materials, energy, and manpower. From the standpoint of

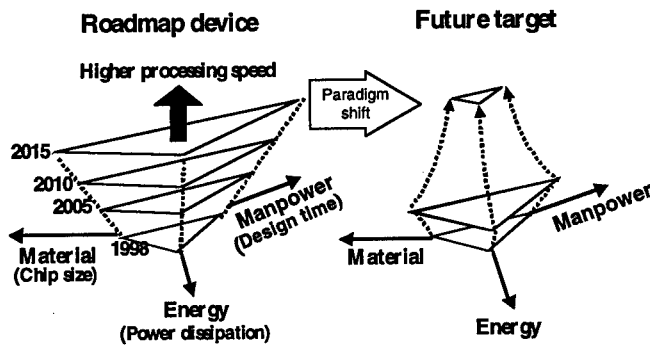


Figure 6. Paradigm shift towards less resource consumption in device development.

environmental protection, it will be difficult to continue on the present path of LSI development. We need a paradigm shift, illustrated in Fig. 6, to devices that require fewer resources and less manpower to fabricate. The increase in chip size is the dominant reason for increased materials consumption. A large chip needs a large amount of fabrication process materials, like chemicals and water. It also requires bigger Si wafers that require larger process equipment. LSI power dissipation is also becoming a factor in terms of energy consumption and environmental pollution, such as the release of CO_2 into the atmosphere. The present Internet access structure presupposes systems working 24 hours a day. The air-conditioned communication hubs that do the switching require huge amounts of electricity. Personal computers and servers are also on for many hours a day, even when they are simply waiting for information input. This type of waste should be minimized by developing new systems with nonvolatile high speed memories. A very high speed data transfer system with extremely high bit rates would also effectively reduce the operation time. The more complex the system, the more manpower is needed to design and debug it. Computer aided design at the higher levels of system design should alleviate this huge manpower consumption. The ultimate aim should be automatic software and hardware co-design system synthesis.

5. Conclusions

This article summarizes the historical driving factors and breakthroughs in the semiconductor industry. We find that the market characteristics are quite different in Si and III-V devices, with the former evolving towards integration engineering and the latter towards bandgap engineering. In integration engineering, advanced technology is driven by commercial DRAM and MPU circuitry. In bandgap

engineering, basic physics and sophistication of epitaxial technology are key, spanning the entire range from basic device concepts to commercial devices. It is interesting that some of the advanced research in Si and III-V devices is converging simultaneously towards quantum dot systems. While it would be premature to conclude that future information processing will be based on quantum computing, it is certain that the smallest devices will be affected by quantum confinement and other quantum effects. From the viewpoint of effective or recycled usage of resources, less material, less power dissipation and less manpower will be major issues in the 21st century semiconductor industry.

On the Edge of Ubiquitous Computing

J. Daniel Janowski and Trey Smith

Compaq Computer Corporation, P.O. Box 692000, M. S. 110605, Houston, TX 77269-2000, U.S.A.

1. Introduction

The computer industry is notorious for dramatic changes and turnabouts. The technology is developing so fast; the money being invested, made, and lost so enormous; and the market so fickle, that it is often difficult to predict who will even survive, let alone who will prosper. This uncertainty exists for even the largest companies. No one is immune. Unless a company keeps evolving and reinventing itself, it will not last long under the hooves of the cattle stampede that is the computer industry. The cattle stampede metaphor is very appropriate, because just like a startled herd in the late-night cowboy Western, the direction it will run once lightning strikes cannot really be predicted. But once it gets going, look out!

One of the most fascinating aspects of technical innovation and market change is that many times it happens so quickly and so transparently, that afterward you don't even remember the transition. In fact, you cannot imagine things ever being any other way.

The landscape is replete with examples. You need look no further than microwave ovens. You would also be hard-pressed to find a store still selling vinyl records. Of course, 8 mm movie cameras have nearly disappeared, courtesy of the camcorder. Remember "rabbit ear" antennas? Who could live without cable TV or a satellite dish today? The Internet — who knew? Artifacts of paradigm shifts are all around us. And more of them are coming. On the horizon are advanced global positioning (GPS) tracking systems and more wireless communications for every vehicle. Are you up for a real shock? DVD RAM and DVD recorders are just around the corner. That means your CDs and CD-ROMs may become obsolete.

A true paradigm shift can be terribly disruptive. Whole markets, industries, and companies can appear or disappear almost overnight. That is why marketing people and company executives are so fixated on paradigm shifts, and why such shifts are so important to identify and anticipate. If you or your company is in the midst of a technology transition, you could be quickly destroyed if you are not careful. Company executives and industry leaders also can end up looking incredibly brilliant, or incredibly foolish — and make or lose fortunes in the process.

Ken Olsen, a true pioneer, brilliant engineer, gutsy innovator, and co-inventor of the mini-computer (a paradigm shift in its own right) refused to believe anyone would ever want to own a "personal computer". Andy Grove and Bill Gates did. The rest, as you know, is history. So, anticipating these disruptions takes up an inordinately large amount of time and is cause for great consternation. But the driving force underlying these transitions in the computer industry actually can be explained by a very simple phenomenon.

2. Moore's Law

In 1967, Intel's co-founder Gordon Moore postulated that the number of transistors on a chip would double every 18 months, and that the relative cost of that improvement would decline 35% per year. While that statement did not appear particularly prophetic at the time, it had then, and still has now, truly astounding implications. First, it says that the increase in processing power and the decrease in cost is exponential. Though the shifts may be modest initially, the changes eventually become massive. Secondly, it says that eventually processing power will become so cheap and fast, the "computer" would become virtually free. And along with these changes, ever more complicated processes would be possible — in other words, the "computer" will be able to do more and more things, and become more and more capable, while it becomes less and less expensive.

Moore's postulate became Moore's Law. Need proof? Consider this: given the millions of transistors on each chip today, and the millions of chips produced each year, it is now estimated that there are more transistors being produced each year than there are raindrops that fall on all the state of California — *and that number is still increasing exponentially!* If you overlay this principle on the computer industry and extrapolate, you start to see some of what lies ahead. And, at the risk of appearing either incredibly brilliant or incredibly foolish, we can start to see the next paradigm shift. But, to do these extrapolations, we need to use some real life examples.

The first real computer that we would recognize as such today (the computer equivalent of *Australopithecus*), was the Eniac developed in 1946 at the University of Pennsylvania by John Mauchly and Presper Eckert. Eniac weighed 30 tons, bristled with 18,000 vacuum tubes, 3,000 blinking lights and used enough electricity to power a small campus. Sensing an opportunity, companies like IBM developed the technology into marketable products. By doing so, they revolutionized the way businesses operated. The computer took its first baby steps and moved out of the laboratory. In a curious footnote, and forewarning of things to come, a market study conducted then suggested that the total worldwide market for computers would never be greater than a dozen or so. That is a good example of what we mean by the potential for looking "incredibly foolish".

3. Out-of-box experiences

So what happened? In the early 1960s, the mainframe moved again out of the self-contained, air-conditioned laboratory, to the back room at the office, and became a "mini-computer". In today's jargon, it moved "out-of-the-box". A dozen years later, in 1971, Intel developed the first microprocessor. These little devices, smaller than a thumbnail, had as much computing power as did the whole Eniac system. By making them smaller and less costly, Intel, Motorola, and others made microprocessors a commodity and empowered people like Steve Jobs and Steve Wozniak to create the personal computer.

Now the "computer" moved from the back office to the office itself – another out-of-the-box transition. Later companies like Compaq developed transportable versions of these personal computers and moved them out of the office altogether. Today the transition is to pocket PCs. Where to next? Well if you follow through with the trend, the computer would now go everywhere – in other words, it will take another step "out-of-the-box" and become ubiquitous.

What do we mean by ubiquitous and how does that start a stampede? Let us look at another similar technical revolution and its own out-of-the-box lessons.

4. Disappearing sparks

In 1752 Benjamin Franklin discovered electricity. It wasn't for some time that electricity would be put to use. But, eventually electric motors and electric lights were developed. At first, light bulbs and electric motors were expensive, specialized devices. Only a few existed, and, being expensive and unreliable, they were used only in limited situations. A visitor to a modern, progressive, early 20th-century factory floor might be taken around to see the one massive electric motor, which, through conveyor belts and pulleys drove the entire factory and most of the machinery in it. The electric motor was on the early 20th century equivalent of the raised floor. But, the electrical industry changed, and its evolution is a harbinger of things to come in the computer industry.

Electric motors became smaller, cheaper and more powerful. Rather than one large motor powering everything, small motors were used to power individual devices (they became personal, just as would the computer). Eventually, the electric motor became tiny, cheap, and simple to use. It then literally disappeared from view. Now electric motors, those massive, very visible "marvels" of a bygone era, are all pervasive, embedded in millions of different devices – tiny, cheap, and invisible.

The same is happening to computers. The personal computer today is giving way to smaller, faster, less expensive, smarter machines. In a further extrapolation of Moore's Law, these ever more powerful and cheaper devices will become essentially free, and as such will find themselves being applied in an ever wider assortment of devices, to do more and more tasks. Like the electric motor

vanishing into the woodwork, the computer will make the final step out-of-the-box, and vanish into the background.

This is not to say that personal computers will disappear altogether. There are still large electric motors around, and personal computers will remain as well, though certainly not looking like the device we see today. But, a tremendous number of new computers — in fact the vast bulk of them — will become invisible, embedded in a wide variety of electrical devices, co-existing all round us. Every electric appliance, whether powered by a wall outlet, the sun, or by batteries, will have an embedded computer in it. And, they will all be communicating with one another.

From a practical standpoint what does that mean to companies like Compaq, or other vendors of electrical appliances; and what impact will this have on the user? Again, let us look at some practical real life examples.

5. The Internet *et al.*

One of the most dramatic changes we have gone through has been the development of the Internet and how it is being used. You know the Internet. Five years ago, the most engaging part of the Internet, the World-Wide-Web, didn't even exist. The Internet was an esoteric lab communication network for Ph.D.'s and programmers. Today the web has 50 million users, with over half of those logging on in the United States alone. Expect 175 million users by 2001, and 1 billion (that's $1/6^{\text{th}}$ of the entire population of the planet) by 2010.

With this increase in the number of users has come a corresponding increase in demand for Internet hosts and servers. There has already been an exponential increase in the number of Internet servers now on the Web. There are over 30 million of them, now. This market represents an incredible opportunity for hardware vendors.

Likewise all that information will mean faster and better methods will be needed to access the information. AltaVista defined the whole concept of search engines. AltaVista now provides instantaneous access to more than 100,000,000 pages of information for more than 16,000 Usenet groups. More than 18,000,000 users access AltaVista each month, and conduct 32,000,000 searches viewing more than 960,000,000 pages, in 25 different languages. That's a lot of 0's — which translates into a lot of potential \$'s.

Search engines will continue to expand and improve their capabilities, not only to keep up with demand, but also to make the Internet easier to use and more fun. Some of that massive exponentially growing processing power will be diverted to improving search engine capabilities, interfaces, and content.

Exponential improvements in the Internet servers and search engines will not by themselves solve all the problems. Other roadblocks will crop up. The exponentially increasing amount of information available and the exponentially increasing number of users mean that exponentially increasing bandwidth will be needed. Fortunately, Moore's Law applies here as well. Compaq is working with

Intel, Microsoft, and other leading telecommunications companies to expand the amount of bandwidth available; to both enable more access, and improve the access that already exists.

A simplified version of the asymmetric digital subscriber line (ADSL) standard is in the works. ADSL will provide 26 times the bandwidth that is possible today on the fastest 56 kbps consumer modems. This increase means more access and faster access. Next up are cable modems and full ADSL.

6. Computing on the road

Ubiquitous computing also means ubiquitous Internet connectivity. Connectivity today generally means in your office or home. But, that will not be the case tomorrow. Computers have already started to show up in automobiles; and fleet management networks using sophisticated GPS tracking systems have become standard issue on many of the trucks driving right next to you today.

The average American spends 73 minutes a day in a car. Today that traveler may have a cellular phone, or GPS tracking system. Tomorrow he will have an integrated computer with its own Internet web address, which will continuously track, download and upload information, and interact with him in the car. Today the user has to be hands-on to run a computer. The embedded, hands-off, ubiquitous, mobile computer of the near future will be much more interactive and automatic.

Voice recognition software will allow the computer and the operator to "communicate" without the need of a keyboard. Integrated GPS, navigation, emergency traffic monitoring, wireless audio and video, communications, entertainment, and Web access will provide the traveler with a variety of new information sources. In many instances these sources will operate automatically. For example, sensing a pending engine failure, or a low tire pressure, the onboard computer would warn the driver, suggest appropriate action, identify a nearby repair shop, provide directions and help navigate there, even call ahead to inform the shop you're coming, and order up a new tire, of the right brand, size and type.

You would even be able to get e-mail messages read to you over your car-phone. And, in conjunction with the GPS system and the Internet access, the car could keep the traveler posted of services in the immediate vicinity, like restaurants, and shops. The possibilities and potential are enormous.

7. Computing in the home

Home computing itself will not stand still. Yes, there will still be the familiar personal computer box sitting on the desk next to the books and printer. But think of the possibilities when the computer starts communicating with all of the other embedded computers around the house.

In the living room will be a large screen TV. But rental tapes and broadcast time schedules will no longer be needed. A user will be able to watch what he wants when he wants, in high-resolution, stereo, surround sound, digital audio and video. Intelligent embedded computers will provide background information. The Web will allow instantaneous access to programming, station notes, and news. Interactive software will allow you to pay your bills, cruise on-line shopping channels, order and pay for products, play games, see and talk with grandma in Cincinnati, send flowers and mail, hold meetings, vote. The world will become suddenly more open and accessible.

Speech recognition software will make doing these things easy. Even the toaster oven may have its own embedded computer. Smart packaging will tell it what to heat, how long to heat it, what temperature it should reach — even adjust some of the cooking parameters to suit your individual taste. The coffeepot will turn itself on, the alarm clock will be set to your appointments, the lights will turn on and off as people move around the house, rooms will be heated and cooled depending on who is in them.

All of this is possible now to a degree. But, cheap, embedded, powerful, ubiquitous computers will make the dream a reality, and much sooner than you think.

8. Conclusion

Cattle stampede or not, disruption or transition, paradigm shift or market trend, the change will happen so fast you probably will never notice it happening. Just like many of the other revolutions in the industry, one day you'll be typing at your PC, and the next you'll be talking to your toaster. And, you won't even remember the time when toasters weren't listening.

Now, all we have to do is figure out the next paradigm after that one. Mr. Coffee, brew up another cup of dark espresso. It's going to be a long day!

On the Future Organization of Hybrid Chip Manufacturing

Eugene A. Feinberg

W. A. Harriman School for Management and Policy, SUNY at Stony Brook, Stony Brook, NY 11794-3775, U.S.A.

Serge Luryi

Dept. of Electrical and Computer Engineering, SUNY at Stony Brook, Stony Brook, NY 11794-2350, U.S.A.

1. Introduction: some trends in the VLSI industry

There are numerous indications that we are at a turning point in the evolution of the giant VLSI industry. For many years the celebrated silicon technology has known a virtually one-dimensional path of development: reducing the minimal size of lithographic features. There is now widespread recognition that this path has brought us to the point of diminishing return. The often quoted Moore's Law — supposed to express the exponential nature of the VLSI progress — is in fact slowing down.¹ In 1965 when Intel's founder Gordon Moore proclaimed his exponential law, the time constant in that exponent — corresponding to the doubling of the number of transistors on a manufactured chip — was once every 12 months. That would be growing by a factor of 1000 every decade. By the mid 1970s when Moore's Law was firmly entrenched, the actual time constant was about 18 months — and this corresponds to a factor of about 100 every decade. By the end of 1980s this was no longer valid and the actual time constant was about 2 years. The 1994 SIA Roadmap assumes a growth of only about a factor of 10 between 1997 and 2007 for microprocessors, implying a time constant of 3 years (to be sure, this projection is likely to be further adjusted).

Appreciation of this hard reality by the semiconductor industry has led to a noticeable shift of investment from new technologies to software and circuit design within existing technologies. There is certainly no shortage of opinion about these trends. Some, haunted by the specter of the steel industry, believe that the semiconductor industry has matured and the research game is over. Others believe the progress in hardware technology will come roaring back, based on innovative research. We certainly belong to the latter category. However, the innovative research that we anticipate will be markedly different from that we have been witnessing over the past 20–30 years. Instead of shrinking the dimensions of Si devices or perfecting exotic compound semiconductor technologies, successful researchers will broker marriages between these technologies. There is no doubt that silicon will remain the dominant

semiconductor material; teaching new trick to the old dog will be the key to success. In this context, we believe that most significant applications of compound semiconductor electronics will be associated with its use in silicon electronics.

With the above considerations in mind, we focus our consideration on the integration of high-performance electronic and optoelectronic devices and small systems with Si circuits, based on advanced packaging concepts and interconnect technology. It is our vision that future electronic systems will have critical needs in on-chip transformation of the signal power among electrical, optical, and microwave media. Communication between relatively small subsystems on the same chip and interchip communications from the chip interior will enable qualitatively new systems.

This point of view has been a common theme echoed by a number of panelists at "Future Trends in Microelectronics" workshops,² and it has been endorsed by other think tanks as well. The U.S. Army Electronics Strategy Planning Workshop (January, 1995) has identified as "Extremely High Priority" the research thrust in Advanced Electronic and Optoelectronic Materials with the following justification:

"It is anticipated that future Si IC technology will evolve to incorporate ultra-high performance electronic and optoelectronic on-chip elements. These elements will facilitate input-output functions from the chip interior, as well as high-bandwidth intrachip communication. Relatively small Si subsystems (less than 100,000 gates) need to be internally inter-connected using minimum-feature rules. To achieve these goals, research needs to be directed toward: (1) advanced packaging concepts; (2) hybrid devices based on mixed compound semiconductor structures on Si; (3) novel device concepts relating to mixed materials architectures; (4) novel very large scale integrated circuit architectures that take advantage of the wide-band communication between parallel subsystems on a Si chip."

The logic of industrial evolution will motivate new paths for a qualitative improvement of system components, other than the traditional path of a steady reduction in fine-line feature size. It is widely accepted that some, perhaps most of the future systems will be on-chip systems with inclusion of foreign elements (e.g. compound semiconductor) into CMOS chips. This inclusion may be achieved by using special islands, interconnected with the rest of the chip via final metal lines.³ This incorporation is done on a whole wafer scale before dicing the wafer into small chips. Foreign elements included may perform functions otherwise inaccessible to silicon, like emission or efficient absorption of light, or they may provide ultra-high performance above the capability of silicon technology.

One of the promising hybrid technologies is *active packaging*, which is a device fabrication technique intended to implement devices on a foreign platform

that perform better than conventionally fabricated devices on their natural semiconductor substrates.⁴ In active packaging certain essential fabrication steps (lithography, etching, metallization, etc.) are performed *after* the partially processed device or circuit is packaged onto a host platform. This often enables the implementation of structures that cannot be realistically obtained in another way, such as those requiring lithography on *opposite* sides of a thin semiconductor film.

One of the most important goals of active packaging is the combination of dissimilar materials (notably, III-V compound semiconductors) with silicon integrated circuitry (IC) on a single Si substrate.⁵ This goal is now widely recognized as an important research direction in microelectronics and is shared by other emerging technologies, such as those based on *heteroepitaxial* and *thin-film transfer* techniques.⁶

The hybrid chip technology is not intended to replace conventional devices, but rather to complement them. Hybrid devices are in a sense "discrete" as their number on the chip will be relatively small compared to that of ULSI transistors. However, development of hybrid chips will offer much higher system performance and broader functions. This technology is compatible with and extends the current "miniaturization" trend in microelectronics as expressed by Moore's Law. Whether or not the microelectronics industry will follow Moore's Law in the next decade, hybrid chips will become an important step in future developments and one can anticipate the arrival of high-volume production of hybrid chip systems. In this article we consider some of the issues associated with the manufacturing organization of hybrid chips. We also present a simple cost-volume model to identify the economic conditions that will drive the organizational paradigm shift, leading to both higher profits and a high entry barrier to the competition.

2. Structure of production facilities for hybrid chips

The paradigm shift to on-chip integrated hybrid systems calls for a radical rethinking of the entire manufacturing process — not only at an individual factory level, but at an even deeper level involving dynamic interaction of distinct factories.

Implementation of hybrid technologies may be summarized in the following four steps:

- (a) standard VLSI processing of a silicon wafer;
- (b) special purpose islands are left virgin on the Si wafer or perhaps pits are etched in to receive foreign elements;
- (c) foreign elements (possibly prefabricated at least partially) are attached to the surface of the Si wafer at the sites of the specially grown islands;
- (d) post-packaging processing, at a minimum including interconnect by final metallization.

The key question addressed in this work is how should such a production be organized? We shall try to answer this question leaving aside the obvious motivation to perform steps (a) and (b) in different physical locations to avoid contamination. That motivation is not a very strong argument because one technological process can be organized in multiple locations in a multi-echelon production process.

The first basic question is whether foreign semiconductor incorporation should be included into the Si CMOS process or separate facilities for steps (c) and (d) should be established.

We believe several arguments are in favor of the second option. This option implies the coexistence of two kinds of factories: one a large and steady Si CMOS processing and fabrication facility (referred to below as the "Foundry"), the other being a multiplicity of relatively small and entrepreneurial compound semiconductor fabrication and hybrid packaging facilities ("Hybrid Packagers", or simply "Packagers"). These arguments include:

- The Foundry uses a stable and mature technology; introduces changes slowly; uses extremely expensive equipment and facilities; is open to orders from outside via well defined process specifications. Because of the economies of scale, steps (a) and (b) are much cheaper when carried out in this environment.
- Several entrepreneurial Hybrid Packagers provide a pool of orders to the Foundry, thus making the flow of orders more steady, which leads to higher utilization of the Foundry facilities. The effect of such pooling is well-known in stochastic models.⁷
- The product differentiation is effectively delayed past the Foundry, making the Foundry more cost-efficient in managing its resources and inventories.⁸
- Hybrid Packagers are high-risk enterprises, often operating *ad hoc* to produce one specific product. They rise quickly and may change the product on a rapid time scale. Their products are systems-on-a-chip, rather than a commodity chip for numerous other systems. The separation from the Foundry frees the Packagers for high-risk innovations.

The economic role of a Hybrid Packager is actually quite similar to that of a contemporary system design house that gives orders to silicon foundries for custom chips. The distinguishing features of this type of enterprise are as above: risk, innovation, and entrepreneurial proprietary nature of the product.

The second basic question is whether the Foundry and the multiplicity of Packagers should form one economic unit.

We believe the answer is negative. Proprietary aspects and rights to a new product reside with the Packager, while the Foundry performs services available to everybody. In our vision, the Packagers will order new products from the Foundry by using predetermined standards via the Internet. The full utilization of the Foundry capabilities by external users mandates that all technologies used at the Foundry be transparent to the users. The Foundries will be protected against

competitors mainly by two entry barriers:⁹ the economies of scale and the high costs for capital requirements to start a new Foundry.

In contrast, the Packager's activity may be quite opaque and protected by patents and trade secrets. This separation will be cost-effective because it encourages competition between Packagers and invites new Packagers to enter the market. These new Packagers create additional products and generate new business for the Foundry.

The separation between the Foundry and Packagers provides advantages for the Foundry, because new Packagers increase the economies of scale for the Foundry. The existence of the Foundry provides the advantage of relatively low entrance costs for Packagers because they do not have to invest in Si processing technologies and they can focus on hybrid technologies and proprietary systems that use them.

At this point let us clearly recognize the counter arguments in favor of both the Si CMOS processing and the foreign semiconductor incorporation being done within the same economic unit. Arguments that suggest this economic organization is the more efficient one may include the following considerations:

- If the Foundry is separate from Packagers, Foundry competitors have easier entrance to the market because the Foundry technologies contain little or no secrets within the industry;
- Profits per unit of production are much higher for system chips than for commodity chips.

The exact economic conditions (e.g., the demand on hybrid chips, product variety, amounts of investments) that will create incentives to the Foundry to make its technologies known to all users, are not clear and should be investigated. Of the many factors influencing the economic decision, we shall concentrate on the demand for the variety of chips generated by independent Packagers. We believe that independent Packagers driven by their entrepreneurial spirit will create a higher variety of on-chip systems. New products will increase the demand. In the next section we consider a simple mathematical model based on cost-volume analysis. This model demonstrates that if the proliferation of independent Packagers significantly increases the demand, then the structure with a Foundry and independent Packagers becomes more profitable for the VLSI manufacturer. Moreover, it creates a higher entrance barrier for competition compared to the economic regime where the same company produces on-chip devices from the beginning to the end.

3. Cost-volume model

Let there be I types of hybrid on-chip devices, $i = 1, 2, \dots, I$, and let index 0 correspond to the standard VLSI processing. We denote by D_i the demand for a particular type of a product, $i = 1, 2, \dots, I$. For simplicity, we assume there is no waste in hybrid technologies, i.e. all non-defective silicon chips are used to create

on-chip systems and their supply is equal to the demand. This assumptions means that:

$$D_0 = \sum_{i=1}^I D_i.$$

We also assume for the sake of simplicity that each type of device corresponds to a manufacturing facility. In reality, of course, different facilities may produce similar devices and one facility may produce several types of devices. This assumption, however, is not restrictive for our analysis because we can introduce fictitious facilities and fictitious products.

Let r_i and c_i be respectively the revenue and cost per unit for each type of production and let A_i be the fixed costs for facility i , $i = 1, 2, \dots, I$. We assume that $A_i(D_i)$ is a non-decreasing step function such that $A_i(D_i)/D_i$ is decreasing in D_i , $i = 1, 2, \dots, I$. In general, the last condition may not hold, but if the demand is high, the fixed costs can be approximated by a function with this property.

If we consider the scheme where the Foundry and Packagers act as independent businesses, we have that the profit for each participant is:

$$P_i = (r_i - c_i)D_i - A_i(D_i), \quad i = 1, 2, \dots, I.$$

The cost of entry for the Foundry is

$$E_0 = \alpha c_0 D_0 + A_0(D_0)$$

where $\alpha \in [0,1]$ is a fraction of variable costs needed for entry.

One the other hand, in the scheme where one manufacturer of on-chip systems is responsible for both the Si chip production and the packaging, then the number of possible hybrid systems will be significantly smaller. Let us assume that only the first J types of units be manufactured in this case, $J \leq I$, and let the new demand be

$$D'_0 = \sum_{i=1}^J D_i.$$

In this scenario we find that the profit is given by

$$P = (r_0 - c_0)D'_0 + \sum_{i=1}^J (r_i - c_i)D_i - A_0(D'_0) - \sum_{i=1}^J A_i(D_i)$$

and the cost of entry is

$$E = \alpha(c_0 D'_0 + \sum_{i=1}^J c_i D_i) + A_0(D'_0) + \sum_{i=1}^J A_i(D_i).$$

In order to simplify the formulae for P_0 , P , E_0 , and E , we introduce additional assumptions. Let the costs, revenues, and demands be equal for all possible hybrid facilities, i.e. $c_i = c_1$, $A_i = A_1$, $D_i = D_1$, and $r_i = r_1$ for all $i \geq 1$. Then

$$P = (r_0 + r_1 - c_0 - c_1)JD_1 - JA_1(D_1) - A_0(JD_1),$$

$$E = \alpha(c_0 + c_1)JD_1 + JA_1(D_1) + A_0(JD_1),$$

and

$$P_0 = (r_0 - c_0)ID_1 - A_1(ID_1),$$

$$E_0 = \alpha c_0 ID_1 + A_0(ID_1).$$

We observe that $P(I) = (r_0 - c_0) - A_0(JD_1)/I$ is the profit per chip at the standard VLSI processing stage of production. The function $P(I)$ is a nondecreasing function because $A_0(JD_1)/I$ is nonincreasing. The condition that $P(I) < 0$ for all I means that for any possible volume the silicon chip production is not profitable. From a mathematical point of view, it is possible to increase the revenue per chip r_0 and make $P(I) > 0$. If this level r_0 is too high, it means that the Foundry cannot be profitable.

Therefore, it is natural to assume $P(I_0) > 0$ for some $I_0 > 0$. This assumption implies that $\lim_{I \rightarrow 0} P(I) > 0$ and $P(I) > 0$ for all $I \geq I_0$.

We have that $E_0 \geq E$ for $I \geq I_1$, where I_1 is any number such that

$$I_1 \geq \frac{\alpha(c_0 + c_1)JD_1 + JA_1(D_1) + A_0(ID_1)}{\alpha c_0 D_1 + A_0(I_1 D_1)/I_1}.$$

For example, we can select

$$I_1 = \frac{\alpha(c_0 + c_1)JD_1 + JA_1(D_1) + A_0(ID_1)}{\alpha c_0 D_1} < \infty.$$

Let $I \geq I_0$. Then $I_0 \geq P$ for $I_0 \geq I_2$, where I_2 is any number such that $I_2 \geq I_0$ and

$$I_2 \geq \frac{\max\{0, (r_0 + r_1 - c_0 - c_1)JD_1 - JA_1(D_1) - A_0(JD_1)\}}{P(I_2)}.$$

We observe that this finite number I_2 exists. Indeed, if the numerator in the last expression is 0 then we can set $I_2 = I_0$. Otherwise, we consider the function $F(I) = IP(I)$. We have that $F(I) \rightarrow \infty$ as $I \rightarrow \infty$. Thus, the inequality holds if I_2 is large enough.

Therefore, if $I \geq \max\{I_0, I_1, I_2\}$, the structure with a Foundry and independent Packagers provides higher profits for Si chip manufacturers and higher entrance barriers for its competitors.

4. Conclusions

We have considered recent trends in semiconductor chip manufacturing. In agreement with a number of forecasters, we believe these trends suggest that the microelectronics industry is at the threshold of revolutionary changes. One of the anticipated elements of the new economic order will be the transition to high-volume production of hybrid on-chip systems. This change will result in the creation of new entities in the semiconductor industry, dubbed here the hybrid packagers, whose role is somewhat reminiscent of the present-day system houses

involved in proprietary chip design. We have analyzed two distinct ways these new entities can be organized: one when they become parts of a larger economic unit involving the mature silicon production, the other when they form a multiplicity of independent entrepreneurial units interacting with a large Si Foundry that produces custom units on order based on common technology. We introduced a simple mathematical model that identifies economic conditions when the second scheme provides higher profits to the Foundry and higher entry barriers to its competitors.

5. Acknowledgments

The authors thank Christopher L. Tucci for interesting comments. Research of the first author was partially supported by the NSF (DMI-9500746).

References

1. See, for example, A. E. Brenner, "Moore's Law," *Science* **275**, 1551 (1997).
2. For some industry opinions, see S. Luryi, J. M. Xu, and A. Zaslavsky, eds., *Future Trends in Microelectronics: Reflections on the Road to Nanotechnology*, NATO ASI series Vol. E323, Dordrecht: Kluwer, 1996; more recent appraisals are available elsewhere in this book.
3. A. E. Kaloyeros, S. Luryi, J. G. Ryan, and J. J. Sullivan, "High performance interconnects for on-chip device integration," *Semicond. International* **20** (12), 115 (1997).
4. S. Luryi, "Active packaging: a new fabrication principle for high performance devices and systems," in: S. Luryi, J. M. Xu, and A. Zaslavsky, eds., *Future Trends in Microelectronics: Reflections on the Road to Nanotechnology*, NATO ASI series Vol. E323, Dordrecht: Kluwer, 1996, pp. 35-43; U.S. patent #5,496,743 (filed 1993, issued 1996).
5. S. Luryi and S. M. Sze, "Possible device applications of silicon molecular beam epitaxy," in: E. Kasper and J. C. Bean, eds., *Silicon Molecular Beam Epitaxy*, Vol. 1, Boca Raton, FL: CRC Press, 1988, pp. 181-240.
6. J. Deboeck and G. Borghs, "III-V on Si — heteroepitaxy versus lift-off techniques," *J. Crystal Growth* **127**, 85 (1993).
7. R. W. Wolff, *Stochastic Modeling and the Theory of Queues*, Englewood Cliffs, NJ: Prentice-Hall, 1989.
8. H. L. Lee and C. S. Tang, "Modeling the cost benefits of delayed product differentiation," *Management Science* **43**, 40 (1997).
9. M. E. Porter, *Competitive Strategy*, New York: The Free Press, 1980.

The End of Scaling: Disruption from Below

Don Monroe

Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 U.S.A.

1. Introduction

The continued scaling of Si CMOS (complementary metal-oxide semiconductor) technology to ever-smaller dimensions faces many critical issues. It is widely believed that these daunting challenges must eventually cause further scaling to slow, and then stop. We discuss an alternative scenario, in which scaling stops not when it becomes *impossible*, but when it is increasingly *irrelevant* to the needs of the mainstream semiconductor market. This scenario has profound implications for warning signs of the end of scaling, and strategies to weather it.

2. Transistor scaling

According to the National Technology Roadmap for Semiconductors,¹ which embodies the collective wisdom of the U.S. semiconductor industry, the transistors in integrated circuits (ICs) of 2012 may be rather similar to those available today, except for size. The gate length of isolated transistors will be about 35 nm, with one-standard-deviation control of about 1 nm, across a wafer as well as between different lots. The gate dielectric will have a capacitance equivalent to less than 1 nm of SiO₂. Junction depths will be 10–20 nm, while maintaining high enough conductivity to avoid impeding the high current drive of these short-channel transistors. To maintain low off-currents and to control short-channel effects, the total thickness of the active area, from the gate to an underlying equipotential such as a second gate or a highly-doped layer, will have to be less than about 30 nm. These well-known constraints pose significant challenges to continuing historical trends of device scaling. Indeed, a major purpose of the roadmap is to communicate these challenges to the academic and equipment communities well in advance, since devising and implementing solutions will require many years.

These long-recognized difficulties posed by continued scaling have engendered numerous predictions of impending saturation of Moore's Law. Many of these predictions have already proven premature; others still lie in the future. Nonetheless, it seems clear that within 10–20 years, scaling will hit some fundamental limits, for example, when the gate dielectrics fall below one atomic layer in thickness. At that future point, scaling will cease. Many proponents of alternative technologies (III-V's, SiGe CMOS, single-electron transistors, quantum

computers, DNA computers, ...) have historically regarded that point as their opportunity to grab the torch and carry it forward into the 21st century.

In this paper, we describe an alternative scenario, with important implications for the timing and the harbingers of the end of scaling. This scenario suggests very different strategies for negotiating the transition period, as well as the attributes of successor technologies to scaled Si CMOS. In the transition we envisage, the benefits of aggressive scaling of Si become secondary to other attributes, eventually rendering further scaling *irrelevant* to the mainstream semiconductor market. Naturally, deprived of its historical economic drivers, scaling beyond that point would radically slow as well.

3. Disruptive technologies

Our description draws on the work of Clayton Christensen, as summarized in his book *The Innovator's Dilemma*.² Christensen has analyzed several industries in which leading firms failed to navigate a significant technological shift. In each case, failure resulted not from management failure or the inability to meet the demands of rapid technological improvement. On the contrary, the firms did an *excellent* job of understanding the future needs of their most important and profitable customers, sometimes expending enormous effort to make the requisite technological advances. Rather, the firms all experienced *failures from below*, in which technologies that performed *poorly* according to the traditional metrics of the industry eventually improved enough to satisfy customers' needs and win the markets.

The best-documented example of this disruptive technology is in the hard disk-drive industry,³ as illustrated in Fig. 1. Within each market segment, the demand for storage capacity, which the disk-drive manufacturers considered their prime competitive attribute, grew exponentially. Nonetheless, the manufacturers were able to improve capacity even *faster*. They did not consider physically smaller drives serious competition, because these drives offered much less storage capacity.

The insurgents, with smaller drives, *did* find eager customers in smaller, emerging technologies that cared less about raw capacity (beyond some minimum) than about other attributes (power, weight, size, *etc.*) in which the new, small drives excelled. Subsequent capacity improvements (faster than required by either market segment) eventually allowed them to challenge the big drive makers. The former market leaders proved largely unable to meet the challenge from their leaner competitors and were *never* able to compete successfully in the new markets. Astonishingly, this transition occurred several times in just sixteen years. Most observers from the semiconductor industry initially regard the disruptive technology model as irrelevant to *our* industry, with its insatiable appetite for bandwidth, processor speed, memory, and so on. The idea of technology improving faster than the market demands or can absorb therefore seems foreign. *Our* environment feels more like that of Lewis Carroll's Red Queen, running as

fast as we can simply to stay in the same place. This perception may be dangerous, however. After all, hard drives compete in very much the same "insatiable" environment! Moreover, as illustrated in Fig. 2 there are some indications of performance overshoot in the semiconductor industry as well.

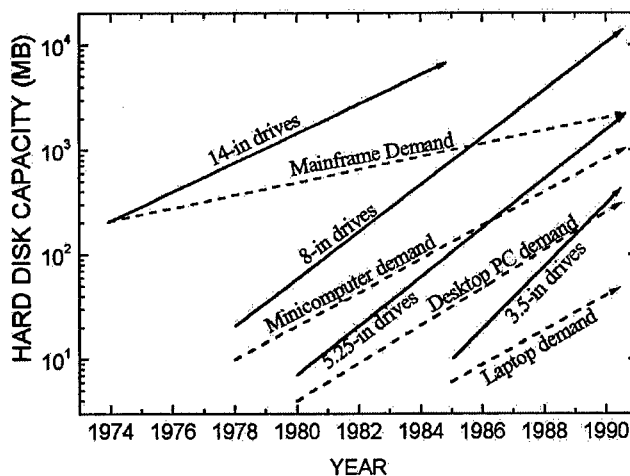


Figure 1. Disruption in the hard-disk-drive industry. Solid lines represent capacity available for various technologies, while dashed lines represent the capacity actual used by various market segments (adapted from Ref. 2).

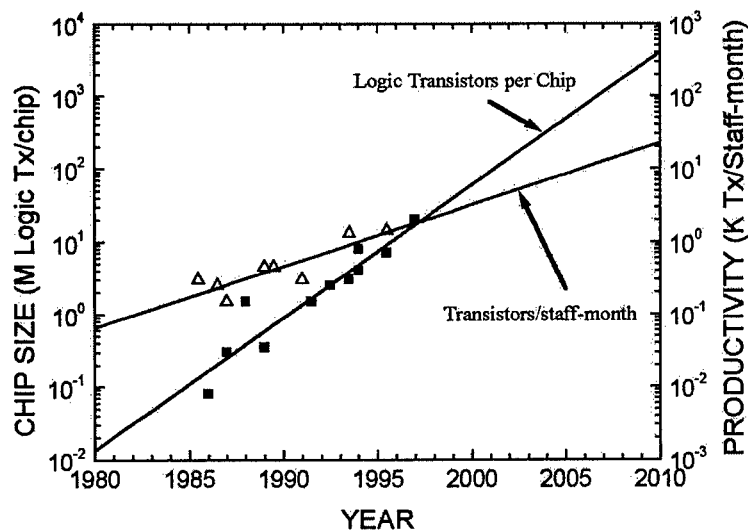


Figure 2. Incipient disruption in the IC industry? (After Greg Ledenbach, Intel assignee at Sematech).

This plot shows rapid improvement (due largely to improvements in CAD tools) in designers' ability to produce reliable logic circuits incorporating ever more transistors. Nonetheless, driven by scaling, these improvements are clearly outpaced by the factories' ability to produce more transistors on a chip. If the designers are the "customers" of the chipmakers, those customers are finding it difficult to absorb further scaling improvements. According to the disruptive technology scenario, scaling is resulting in *performance overshoot* in a historically critical attribute of ICs: transistor count.

Of course, the precise date of "overshoot" depends on one's market segment. Comparing the left and right axes, the plot shows a crossover in 1997 for a design effort of 10,000 staff-months/chip. This scale of effort may be tolerable for a microprocessor powerhouse. For other segments that sell chips more cheaply and fewer of each kind, the design costs of maximum-size chips are already excessive.

4. Microprocessor trends

Microprocessors and personal computers have been a primary driver of the microelectronics industry. Since they are also a segment where many of us have daily personal experience, their dramatic evolution has a visceral reality.

- *Software/hardware co-evolution*

Any PC user cannot help being frustrated by the recurrent obsolescence of computers that had recently seemed entirely adequate. Many regard the huge demand for memory, processor speed, *etc.* of the latest software as incommensurate with the modest improvements in "productivity" they make possible. For example, the fact that a particular toolbar bounces convincingly and makes an amusing whooshing sound when clicked does not seem like a productivity improvement.

Naturally, power enhancements *have* enabled such features as the ability to deal gracefully with images and sound on the PC. An objective observer will also recognize the poor quality of many of these representations, and thus the potential for even more processing capacity to improve them, especially if one desires full-motion, high-resolution video on the desktop. Expanded processing power also takes pressure off limited bandwidth resources, allowing real-time compression and decompression of data streams. The computer and software manufacturers have a clear need to convince ordinary consumers of the need for three-dimensional interactive games, real-time image rendering, and other processor intensive applications. Satisfying such voracious increases in demand is the only way to continually sell more powerful processors.

- *"Segment zero" personal computers*

The recent emergence of low-cost (under US\$ 1000) personal computers is a clear example of disruption, albeit at a higher level than the chip technology. As expected for a disruptive technology, Intel discounted the significance of this

market segment early on (although its hungrier competitors did not). Only after the explosive growth of this segment did Intel introduce lower-performance processors specifically targeting this market. These developments do not themselves presage the end of scaling, since even the down-market processors incorporate up-to-date scaled CMOS. What this example *does* show is that customers (especially in the consumer-electronics market) will sometimes forgo steadily improving performance in favor of other attributes, including cost.

5. Implications for R&D strategies

Together with anecdotal information, there are some signs of performance overshoot in the less demanding tiers of the IC business. In the rest of this article we assume that this overshoot implies the industry is vulnerable to disruptive transition in the near future and explore some of the implications.

- *"On-the-roadmap" research*

The SIA Roadmap publicizes serious challenges facing continued scaling. These challenges include major equipment and infrastructure issues, such as lithography, as well as device and processing questions such as shallow, low-resistance source-drain extensions. The roadmap gives assurance to researchers

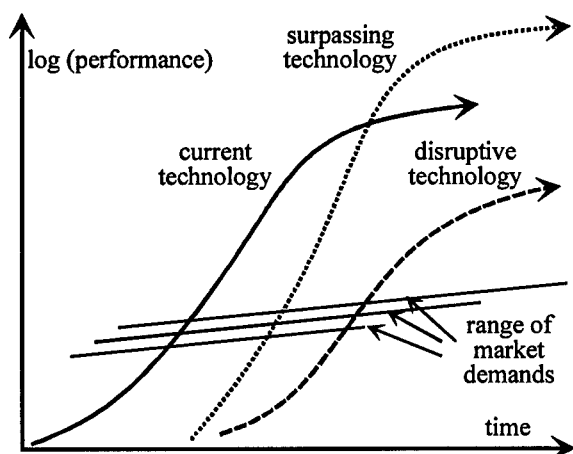


Figure 3. Contrast between sustaining and disruptive technology advances. Advances in the current technology (bold solid line) must eventually slow down, allowing a superior but less mature technology (dotted line) to surpass it (given sufficient development resources). Such a transition can *sustain* the advance of technology as traditionally measured. However, once the current technology exceeds market demands, a disruptive technology (dashed line) providing *other* advantages may beat out the current technology, even if it is inferior according to traditional metrics.

working on narrower questions that their research will be relevant in the future, and gives them the confidence to explore the difficult, detailed issues that confront proposed solutions. It should be obvious that this confidence is misplaced for technology generations beyond a disruptive transition. In light of the uncertainties of *all* true research, however, the continued focus on issues highlighted as potential showstoppers or necessary innovations for the Roadmap remains a good bet for research. The innovations may not find application as soon as expected in the Roadmap or in the same form, however.

- *Near-term development*

The challenges for nearer-term projects are more serious, since they require substantially greater investments, whose timely recovery is critical to profits. Chipmakers, and to an even greater degree equipment makers, are put at substantial risk by the possibility of a disruptive transition. Interestingly, the very success of the roadmap process will make any such transition especially traumatic, encouraging unwarranted confidence in the momentum of continued scaling as long as it is technically possible. Any reduction in the pace of scaling or the market demand will have devastating financial consequences to companies trying to ride the leading edge.

- *"Off-the-roadmap" research*

A large segment of the academic community performs research advertised as technological, but not geared toward extending scaling. Some of this work may enable novel architectures or processing techniques. For example, the architectural principles underlying biological information processing make effective use of parallelism to provide rapid processing with relatively slow devices. Understanding and applying such principles even in relatively antiquated Si technology could improve system performance and power consumption by many orders of magnitude. Research in such areas, while speculative, provides real potential for major breakthroughs.

- *"Beyond-the-roadmap" research*

In contrast, significant effort aims to provide "beyond-the-roadmap" functionality. The goal is to extend the performance, defined according to the *traditional* metrics, beyond an expected saturation of CMOS capability, as illustrated by "surpassing technology" in Fig. 3. The classic example is III-V electronics, but equally relevant might be single-electron memories, thought to extend memories into the Terabit/cm² regime. It is axiomatic in the disruptive technology paradigm that such technologies are addressing the *wrong question*. Well before CMOS has encountered physical limits to scaling, it will have overshoot the necessary performance as traditionally measured, and new differentiating attributes will become important. This is perhaps the most profound conclusion of the scenario described in this paper: "beyond-the-roadmap" research *is unlikely to be directly useful*.

6. Economic drivers

The performance improvements engendered by aggressive CMOS scaling are well known. They include both direct effects on the transistors — improved current drive and reduced capacitance of scaled transistors — and indirect effects of reduced area — reduced interconnect delays and power for smaller chips. The performance advantages of scaling Si have enabled awesome improvements in CMOS performance, while keeping power consumption modest. However, performance improvements alone would *not* have justified the massive R & D investments in scaling. The fundamental driver has been reduced *cost*.

Interestingly, the cost of processed Si per unit area has not changed much: it has remained around \$5/cm² for several decades. The cost that *has* been following Moore's Law of exponential decrease is the cost per transistor. More commonly, this is expressed as "cost per function," that is, the total chip cost divided by the number of logic gates, or some similar function. Chip cost, however, has *not* decreased by many orders of magnitude. Rather, customers have been convinced to buy ever *more* function for roughly the same cost. For example, Marty Lepselter observed⁴ that the current generation of DRAM always costs about $\$ \pi$.

Once entire systems on a chip are available, the economics will change. Does it make sense to buy ten systems on a chip? At this point, the relevant quantity will not be "cost per function," but "cost for *my* function". If customers stop desiring ever-greater functionality, continued scaling makes economic sense only if it reduces the cost for a fixed chip size. Disk drives offers a sobering example of this phenomenon: for *every one* of the new generations shown in Fig. 1 the cost per MB (cost per function) was *higher* for the new drives, when they were introduced. This higher cost did not prevent their eventual dominance.

An essential aspect of the economics of scaling is integrating functionality on a single chip. As with device scaling, this integration includes both performance and economic advantages. These advantages are more complicated, however, when it comes to integrating multiple, disparate device types on a single chip.

Single-chip integration allows increased speed and lower power, as well as more and wider data paths between sections. The speed and power improvements reflect the large parasitics (capacitive and inductive) inherent in traditional packaging technologies. In addition, off-chip connections are subject to uncontrolled terminations and so must accommodate electrostatic discharge, variations in load, *etc.* The required protection and design margin further increase the area. The large metal pads used for contacts may consume precious Si area. Solder connections and metalization on the board remain much larger than those on the chip, and the packaging overhead can be a substantial contributor to board size. Most of these issues, however, are really shortcomings in inter-chip interconnection schemes, rather than intrinsic drivers for single-chip integration.

Indeed, there are many pitfalls in single-chip integration. The well-known difficulty of providing "known-good die" for multi-chip modules is *not* solved by integration. Indeed, reduced pin count per transistor only magnifies the testing challenges for large circuits. Even more challenging is the incompatibility of

process flow for various types of device, for example, CMOS, bipolar, dense and/or programmable memory, and the capacitors, inductors, and resistors useful in analog applications. In some cases, for example BiCMOS or CMOS integration, the devices are so interspersed that developing a compatible process flow is compelling. In other cases, the integration is driven by imperfections of interconnection. Traditional CMOS wafers are also a highly flawed environment for isolating sensitive analog circuits from noisy digital circuits. Thus, for some applications, the economic stimulus for integration may weaken before the entire system is on a single chip, if advanced packaging schemes can minimize these inadequacies of the interconnections.

7. Potential challengers to scaled CMOS

We believe that in the next few years the overall integration level on a chip will cease to be the most important attribute for IC technology. Identifying the vulnerability of Si scaling to disruption is only the beginning, however. The much harder task is identifying candidates for disruptive technologies that will invade the less challenging traditional and emerging markets. The successful candidate will *not* be superior in sheer density of transistors or raw transistor speed, but will offer other attributes, including cost for desired function. Following are a few possibilities. Inevitably, without the luxury of hindsight, this list is arbitrary and almost certainly misses important candidates.

- *Design productivity and re-use*

One response to the design productivity gap illustrated in Fig. 2 is to try to re-use ever-larger blocks of existing designs in new circuits. While this is a traditional goal *within* companies, there is a growing trend toward sharing designs *between* companies (for an appropriate fee). The growing importance of foundries reinforces this trend, as some of them would like to offer not just a process but proven subcircuits as well. It seems clear that designs that function acceptably on a variety of fab lines cannot fully exploit the capabilities of the process. Rather, the goal of optimum performance is traded off for greater flexibility and economy of design resources.

Another well-established route to improve productivity is to re-use the entire microprocessor, and to take advantage of its general-purpose programmability to implement arbitrary functionality in software. This strategy will always be a very effective means to rapid implementation of new algorithms, and provides other advantages such as remote upgradability. To date, however, this strategy has supported, not undercut, increasing microprocessor power, as functionality of dedicated chips is substituted by the underused microprocessor.

- *Integration of diverse functionalities*

The traditional approach to incorporation of alternate functionalities is the inclusion of multiple process modules. At Lucent, this approach is called the

"superchip", in which each process enhancement, such as flash memory or BiCMOS, is totally compatible with the core CMOS process. So far, these enhancements have been achieved without any sacrifice in performance of the fully scaled, core digital CMOS process. However, developing such compatible processes may become harder.

As a possible leading indicator, one company recently introduced a process optimized for the fabrication of CMOS cameras. This process is far from cutting edge in its CMOS performance, but for those interested in camera functionality that may be of little importance.

Another candidate is the functionality provided by micro-electrical-mechanical systems (MEMS). While the return to devices with moving parts seems to conflict with the historical strengths of Si ICs, the reliability of such mass-produced parts as accelerometers for airbags suggests otherwise. Moreover, the fabrication techniques include lithographic patterning and wafer-scale processing, leveraging many of the same economies present for CMOS fabrication. In addition, MEMS can potentially provide small, cheap, electronic interfaces to the non-electronic world, including chemical, mechanical, optical, and magnetic signals. Of course, MEMS processing is not easily integrable with CMOS processing. Existing examples of integrated CMOS/MEMS processes do not approach the state of the art, but even relatively archaic CMOS can substantially enhance the usefulness of MEMS.

Just as there are serious economic and performance drivers for single-chip integration, but there are also serious economic and performance drivers against it. Because many performance problems arise from excessive packaging parasitics, there is continuing interest in packaging that reduces these parasitics. Thus, one can imagine a pre-testable chip-scale package using a low-parasitic, high-density array to contact a compatible board carrying other chips, including area-hungry passive components. Another strategy for improved packaging is optical interconnect, with its potentially high speed and low parasitics. As before, integrating this functionality may not be compatible with ultimate CMOS performance and density, but for many applications it may be more important.

- *Alternative materials systems*

Si can be expected to remain the king of standard semiconductor applications. A novel materials system must include attributes not necessarily quantitatively better but qualitatively distinct from those of Si CMOS.

Organic semiconductors exhibit such attributes. According to traditional metrics, these semiconductors are dreadful, exhibiting mobilities orders of magnitude below those of single-crystal Si, and limited prospects for processing at elevated temperatures. However, they could potentially be printed cheaply on surfaces, including plastics and other flexible materials, and thus used in ways previously unforeseen. The "bendable television" may not seem necessary, but it illustrates the dramatic new possibilities that may become indispensable.

It might appear strange to include III-V electronics, e.g. GaAs, in a list of alternative materials systems, since technology disrupters are generally *lower* in

performance and cost than the reigning technology. Note, however, that in Fig. 3 the challenger lags behind the current leader in the attributes currently valued by the marketplace. This ranking implies that the actual attribute most valued by the mainstream market is *not* speed, but overall integration level (and thus cost), as well as power. Imagine a future in which GaAs (currently a niche player for "low-level" applications not requiring the ultimate in power or integration) continues to improve in integration level. As Si overshoots the integration needs of the marketplace, the door is open for other attributes, including raw speed. Naturally, this performance must not violate other market constraints, including cost, and the future for III-V's in mainstream electronics does not appear terribly bright. We introduce the idea only to emphasize the critical nature of identifying the correct attributes for quantifying "performance."

8. Summary

We have presented here a hint of a future of the semiconductor industry that contrasts sharply with the standard view. In that view, scaling of transistors and interconnect, as envisioned in the SIA Roadmap, continues unabated, providing vast increases in the number of transistors on a chip and similarly huge reductions in cost and power per transistor. At the same time, many companies see the highest value added shifting to systems-on-a-chip: the incorporation of processes in addition to digital logic CMOS, notably high-density and/or programmable memory and full-function analog and rf. The standard view assumes that both scaling and systems-on-a-chip will coexist.

We have suggested that the economic incentive for continued scaling will dissipate as the available level of integration of digital components overshoots the demands of a progressively larger fraction of the market. Silicon will remain the primary materials system for the highly integrated digital functions. The value added will increasingly shift to other functionalities.

At the 1984 International Electron Devices Meeting, Marty Lepselter⁵ likened Si to steel, which remains a primary structural material in spite of niche competition from more exotic materials. It is interesting to note that the steel industry was one of those shown by Christensen to have experienced a disruptive transition in recent decades. That transition did not reflect a dramatic improvement in quality; indeed, the new material was generally inferior. Rather, novel factories ("minimills"), widely adopted in the Far East, progressively invaded the lower, less profitable segments of the steel market. Traditional mills, for example in the U.S., eventually lost most of their market. Perhaps a similar fate awaits the leaders of the IC industry today.

References

1. *The National Technology Roadmap for Semiconductors*, 1997 Edition, Semiconductor Industry Association.
2. Clayton M. Christensen, *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*, Cambridge, MA: Harvard Business School, 1997.
3. Clayton M. Christensen, "The rigid disk drive industry: a history of commercial and technological turbulence," *Business History Rev.* **67**, 531 (1993).
4. M. P. Lepselter and S. M. Sze, "DRAM pricing trends — the pi rule," *IEEE J. Circuits Dev.* **1**, 53 (1985).
5. M. P. Lepselter, "Silicon technology — the new steel," *Tech. Digest IEDM* (1984), p. 9.

Considerations Beyond Moore's Law

P. R. Jay

*Nortel Advanced Technology, P.O. Box 3511, Station 'C', Ottawa, Ontario
Canada, K1Y 4H7*

1. Introduction

A number of articles in the scientific and popular press have recently discussed the potential for changes in the trend established by Moore's Law, namely a continued doubling of transistor count (or more generally, functionality) about every eighteen months. Since its initial proposition in 1965, the same relationship has held true with minor adjustments for more than 30 years, and shows signs of continuing for at least another decade. The predicted rate of development has proved to be so dependable that the technical community actually plans on the basis of that anticipated growth, to the point where it has almost become self-fulfilling. With the massive investment involved in the semiconductor industry, it is not surprising to encounter vigorous debate over "when and why" Moore's Law may cease to be valid.

Curiously, when Mr. Moore said at a recent conference, "Some time in the next several years we get to some fundamental limits",¹ the various reports of this comment interpreted it either as "Mr. Moore claims that his law will continue for many years yet" or as "Mr. Moore suggests that the trend will flatten out in the near future".

Perhaps the most impressive feature of this ongoing dependence has been the way in which many different technological advances have conspired to sustain the progress, particularly in the face of earlier adamant predictions as to its demise. The purpose of this article is not to forecast when and how Moore's Law may break down, but rather to assume that it will some day, and to consider what alternative technology developments might be considered and in particular, what suggested research challenges, if addressed in the next few years, might better position us to evaluate and apply those alternative directions.

It is very likely that the semiconductor industry will continue to make spectacular strides over the next 10–20 years. However, when we imagine an equally tremendous infrastructure springing up to support some alternative technology, it becomes obvious that the response time will be more than just a few years. Hence the value of some creative reflection at this stage.

The discussion is a personal position paper drawing from my own talk at the "Future Trends in Microelectronics" Workshop at Bendor in 1995,² as well as presentations and discussions at other meetings, and in particular from an informal "think-tank" of specialists from various parts of industry and academia, that was held at the Nortel Institute for Telecommunications, University of Toronto in

October 1997.³ The following sections consider various research directions associated with or tangential to this topic, in the hopes that some may spark off ideas for new or modified research projects. More detailed treatments of some of these topics are available elsewhere in this book.

2. Lithography

A powerful force in the continued increase in device density of semiconductor chips is that of progressively more sophisticated lithographic techniques that enable the reliable fabrication of transistors and interconnect to ever smaller dimensions. The recent Semiconductor Industry Association (SIA) Roadmap⁴ suggests that by 2010 transistor gate dimensions could be as small as 0.05 μm . Depending on the effective channel doping levels, this figure means that the number of carriers in an active device channel would be on the order of 10, a small enough number to raise serious questions as to the ability to make identical transistors, and even whether their logic states can be considered as stable.

Clearly demonstrations of "single-electron" transistors have already been reported, so the question is not so much "can such a device be made?" as "how will circuit functions be assembled?" given that the statistics of behavior (and tolerances) of such a device will be substantially different from present day devices relying on many hundreds of carriers per channel. A related question is "if we continue to make transistors at these dimensions, should we still aim to make them as distinct '1' or '0' state devices, or should we perhaps accept their statistical variability and group them in ensembles?"

The capital investment required to perform lithography at these dimensions will be substantial, and since the main driving force for miniaturization has so far been the push for reduced cost (with associated advantages in performance speed, power dissipation etc) the merit of making a smaller device is questionable if it costs substantially more to manufacture.

3. Cost and the lack of competitive differential

So the finer lithography or more dense integration has to offer a basic cost advantage to be acceptable. From a purely industrial viewpoint, the larger wafer sizes and higher capacities of semiconductor processing plants are moving towards a regime where a few large entities (or groups of entities) will control "megafab" facilities in special geographic locations. The rising complexity of highly integrated chips tends towards greater design automation, increasingly high-level design, and (especially as we enter a regime of more non-linear devices) a need to use only the well-characterized libraries of fab-tested elements.

A consequence of this trend will be to limit the scope for innovation through basic design, since virtually all designers will be using the same core components, and designs will be dominated by "Intellectual Property Cores" and inviolate design blocks. At this point it becomes much more difficult to maintain a healthy

market competition and costs will tend to rise again, suggesting that we are presently in a trough on the cost vs. time bathtub curve.

The economic forces involved also reflect the demand of the marketplace, and if this shifts from a situation dominated by professional users requiring state-of-the-art performance, to one where consumers are the largest sector, and for whom price is a very decisive factor, then the notion of cost vs. performance may select out some technically feasible, but overly expensive innovations. This time would be the ideal moment for a radically new technology that was not necessarily faster, not even denser, but much lower cost, to make a useful entry onto the scene.

A challenge here is to predict the demand that will drive the major investment in new silicon fab lines and forecast their depreciation. Even now some major semiconductor manufacturers are addressing issues of excess capacity. When the traditionally exponential growth rate of the semiconductor business starts to ease off, those suppliers who chose to make lateral investments in alternative technologies will have more options.

Indeed, the trend for industry to invest in university research provides an excellent means of complementing the industrial R&D programs with a view on longer range activity, at the same time as addressing the shortfall of funding for the academic research community. This approach, however, carries with it a responsibility on the part of the industrial community to select and encourage projects that attack both near-term needs, and longer-term (more basic) research, lest the "food-chain" of research support ultimately exhaust itself for want of diversity and innovative vision.

4. Functionality inflation

The predictability of the "Moore's Law" trend has led to a situation where the increased functionality that the semiconductor technology fought so hard to achieve becomes rapidly squandered by software developments that consume 2-3 times the memory for each new issue of a given application. The real benefits of the hardware advance should therefore be measured against a functionality scale that is corrected for inflationary effects over time.

This connection suggests that the anticipated growth of chip capability has met a Parkinson's Law effect, where (to paraphrase the original) "the software (originally the 'work') expands to fill the memory (originally the 'time') available". A finite risk exists that when increased device counts become available, then designers add functionality that users do not really want, so that the cost per function may decrease, but *not* the cost per *desired* function. If some movement towards more economic software were effective, then the range (or benefit) of the Moore's Law trend could be substantially extended. In fact, consumer opinion is rapidly becoming aware of the compounded complexity of commercial software in terms of its increasing vulnerability to hidden bugs and incompatibilities, and a more modular and controllable (and hopefully more reliable) mode of software design is coming into effect.

5. Demand for new types of functionality: applications pull

Indeed there is a growing number of cases where new applications are so memory-hungry that it is appropriate to question the suitability of them to the processing medium. Some of the "fuzzy processing" functions that will help to facilitate man-machine interfaces (speech or character recognition, *etc.*) are based on neural network-type programs that essentially use very large numbers of logic gates to simulate a heuristic function. If an alternate device or processing medium could do this more efficiently, then major responsibilities of current systems could be devolved to that medium. However, at this time there seems to be relatively little activity addressing fundamentally different technologies that would evolve a "fuzzy device".

Indeed, it seems ironic (and intuitively questionable) that while we strive to achieve precise '1' or '0' states at an individual device level, large portions of the final circuitry may be given up to far more analog decision-making, even to the point of smoothing out the granularity of the binary decisions at the elemental level.

This type of convergence of new demand around the time of growth saturation of an existing technology creates a major opportunity for a *disruptive technology* that challenges the existing paradigm from an unexpected direction.⁵ This convergence is also an opportunity for the research community, since there are various candidate technologies requiring evaluation. In 1995, at the Bendor meeting, Herb Kroemer described his "Lemma of new technology" as follows:⁶

The principal applications of any sufficiently new and innovative technology always have been — and will continue to be — applications *created* by that technology.

This caution is a valuable one, and when combined with the concept of a disruptive technology it is worth remembering that many new technologies enter the field on the wings of an application that is far from their ultimate domain of success. Nevertheless, an open mind to new application ideas is a valuable driver to technology development.

The current explosion of communication capability is both resolving current needs and creating new ones. While making it easier for individuals to access global reserves of data, the Internet also furnishes more e-mail than can be answered and more information than can be assimilated. Hence, applications which can palliate this burden will find ready acceptance. "Enter stage left" some new intelligent tools that will do contextual data mining and analysis for the user. For effectiveness, these tools might not be most usefully situated at the user extremity of the system, but perhaps at a local node. In an ultimate scenario, the "computing power on the desk" might well decrease, but find more effective and selective ways of using capability at a nearby, or remotely accessible server. It will be interesting to see at which end of this link the more radical new device technologies will have the most impact!

6. Device directions and technology options

The extent to which one type of device (basically a field-effect transistor) has dominated the development of electronics is witness to the elegant simplicity and flexibility of that type of structure, and in particular, the extent to which it can continue to be miniaturized. There are, however, many other types of device function and other materials systems, which offer possibilities that could become the seeds of future major industries.

- *Multi-level logic*

Using conventional devices there have been many demonstrations of logic systems using multiple states, and recent announcements indicate that such technology is now manufacturable and can already offer a useful increase in storage density. More sophisticated device structures (e.g. resonant tunneling transistors) also offer multi-level functions with integrated logic operations using many less components than with a classic field effect device, although manufacturability is not yet confirmed.

- *Nano-electronics*

The pursuit of finer lithography has opened the door to structures that are logical extensions of present device concepts, but which enter a different realm of operation. For example, progress on single-electron devices has demonstrated feasibility of basic logic functions,^{7,8} although significant challenges remain. In particular, the ability of a single-electron device to drive subsequent stages appears to limit the applicability to some type of memory function, and more understanding will be needed to ensure robustness of operation against environmental upsets. Even if ultimately applicable to digital logic architectures, this type of device would represent a new class of technology and require a fundamentally new set of design rules and manufacturability statistics before being commercialized.

Clearly, studies of these new types of devices will benefit from parallel activity addressed at their behavior in large numbers. Interconnect issues then become significant, both in terms of compatibility of physical dimensions and especially regarding appropriate electrical characteristics. Distinctly new solutions to manufacturable (and possibly reconfigurable) interconnects are needed. One-dimensional wires might serve as such interconnects, although these should likely be regarded as devices in their own right, in terms of being an interconnect medium capable of multiple states and strongly resonant transmission characteristics.

- *Interconnect technology*

Interconnect technology has never been as glamorous a field of research as that of device technology, however the scope for fundamental impact is substantial. Even in the domain of passive interconnect, some basic innovations can transform the effectiveness of a given device technology, especially where *RC* delays are becoming dominant at high speeds and small dimensions.

It is interesting to observe that nature rarely indulges in passive interconnect. Most linking functions are also involved in signal regeneration, or even pre-processing en-route to the destination. Another key attribute that still eludes conventional electronics is the ability to reconfigure, either on demand, or autonomously to overcome temporary or permanent loss of function in one area. The discovery of a reconfigurable, compact, low-loss interconnect medium would open profound new possibilities for both conventional technologies and some of the exploratory ones.

Perhaps the most elegant solution will be some integrated medium that can do both signal processing and interconnect within the same adaptable architecture. Since it is likely that the new and existing technologies will need to work together, a part of this challenge becomes finding a new system that will interface compatibly with the traditional signals of our present medium.

- *New materials*

My own personal involvement in this discussion goes back several years to the analysis of future possibilities for GaAs and III-V devices in contrast to the Si family. Whilst the fundamental characteristics of the III-V family of materials (and especially the various heterostructures and opportunities for bandgap engineering) offer significant performance advances, especially at a given lithography, there is no real case to be made that a III-V technology will ultimately displace Si as the economic champion of large scale integration. At a given point in time (and the relative advantages are very time-sensitive since they depend very much on market volume and evolving maturity of a new technology) the relative advantages of one technology over the other usually only represent a factor of 2–3 times, either in cost or performance. This type of differential does not qualify as a truly disruptive technology, even though it must be acknowledged that the advent of GaAs has caused or accelerated significant progress in many aspects of Si technology, faced with the availability of that competition.

In fact it is still probably true that in the limit of small dimensions, the actual material system becomes relatively second-order.⁹ As such, we should be widening our search to new material systems, especially given that, for new classes of consumer application, it may not be necessary to offer a *faster* device, rather one which is substantially (e.g. a factor of 10 or more) cheaper but may still offer *sufficient* performance to win the day.

It is in this context that work on organic alternatives to the conventional semiconductors becomes of significant interest. The major attraction of this type of technology lies in the concept of a fabrication line resembling a printing factory — with large rolls of wide-dimension substrates flowing through huge batch printing operations,¹⁰ and producing devices at very low cost, on flexible and very thin (and very cheap) substrates. Recent work on "plastic transistors"⁵ indicates usable performance for low-speed applications, and the subject definitely merits more attention than it currently receives from a community used to advances in the high speed domain.

The community of organic chemists connecting into the world of semiconductor device specialists is presently quite limited, and deserves to be

encouraged. The subject also presents some scope for valuable simulation work on what could be done with devices with modest mobility, but costing virtually nothing.

- *Self-organized structures*

Although "self-organization" can be defined in a number of ways, my interpretation refers to a technology that achieves some relatively complex structure without the need for detailed external intervention. The trap here may be in looking for mechanisms to achieve self-organization in the regular manner of designs such as we currently produce. This focus might eclipse the broader possibilities of an architecture based less on geometric precision than on "teachable" irregular configurations, possibly taking advantage of reconfigurable interconnect as discussed above.

The coded structure of DNA molecules, coupled with their pre-determined patterns of replication, potentially satisfies the self-organization criterion. Indeed, some studies have suggested using DNA as a carrier/keying medium, so that molecules exposed to a substrate would migrate until they found the DNA-keyed location, at which point (and presumably in a specific orientation) they would deposit. This scheme offers a method of controlled deposition from solution onto a given substrate, after which the DNA could presumably be removed.

- *Optical devices*

The beauty of the III-V semiconductor family is the relative compatibility of optical and electronic device technologies. However, successful co-integration of those functions on to a chip has always encountered the hurdle of the energy lost in making the electrical/optical transition, or vice versa. Few situations lend themselves well to all-optical handling, and so the challenge of an economic, integrable optoelectronic interface remains. For future systems, an interesting area of study is one where a transition is necessary at some point (e.g. a synthetic optical recognition system), where simulations of the partitioning might help to drive studies of the degree to which optical processing can be achieved prior to electronic conversion. Here, too, it is important not to restrict the choices to the input parameters of conventional electronic systems. "Neural-based" technologies, for instance, might offer quite different characteristics, potentially with wider dynamic range. Nature has found some very ingenious solutions to deal with signals from different kinds of sensors.

- *Sensors*

One criterion for a good sensor is usually that of linearity, simply because it is then more easily connected to the next stage of processing. However, biological sensors usually have quite non-linear characteristics compensated by feedback systems of which they are an integrated part. As a result the dynamic range and overload capability are often spectacular. A new domain of sensor applications that could use these capabilities is the automation of our interfaces with technology media (cameras, microphones, chemical detectors, motion/position sensors, *etc.*). As we look ahead (not necessarily forward!) to the increasing

pervasiveness of information/computing technology in our work and home lives, then the scope for broad-ranging sensory detection increases dramatically. Already automotive electronics is developing in that direction, and the home/health care context additionally provides many opportunities for detection of personal or environmental parameters.

Essential to these consumer-focused functions will be low cost, probably meaning the "on-chip" integration of sensors with the associated processing capability.

- *Micro-mechanics*

The impressive achievements of micro-electromechanical systems (MEMS) fabricated using Si technology promise new applications, especially where sensors, transducers, and integrated functions are required. Micro-actuators are of potential interest in a number of medical situations, especially where small size makes a function implantable *in vivo*, or where exceedingly small displacements are involved, such as cochlear implants. There is no fundamental reason why branches of this technology could not develop in other than Si materials, for example III-Vs, where the facility of combination with optical functions could offer micro-optical benches with auto-alignment, self-focusing optical arrays etc.

- *Quantum computing*

The notion of a state machine defined in terms of, say, spin states of particular nuclei in a lattice clearly offers the potential for very densely integrated functionality. At this time it would not seem fair to suggest that such an approach is a serious contender to dethrone conventional Si technology, especially given that: a) reading the states requires equipment many times larger than the sample enclosing the state machine, and b) the stability of these states in relation to energy fluctuations of the local environment would potentially be a major issue. It is, however, tantalizing to suggest that continuing to extrapolate Moore's Law out to 2020 suggests that we then reach the point of storing information in a single quantum state.

7. Design approaches and design system requirements

With hindsight of experience bringing GaAs technology to application, one of the longest development times was the establishment of a design medium with sufficient reliability (i.e. margins based on real data) and familiarity to make it a palatable alternative for the circuit designers involved. Given that the discussion above suggests the need to design in new technologies substantially more foreign in nature, it is even more important to anticipate the time required to develop a design system appropriate to a given new technology.

Especially in the case of a modifiable or adaptable technology, or even just for a technology where it was acknowledged that the individual elements were not necessarily expected to be identical, then a fundamentally new approach to system architecture may be required. This is an excellent study topic that could, in itself,

help to highlight which device approaches lend themselves to effective combination. For example, there has been work on evolutionary design approaches for a hierarchical multi-layer structure for retinal detection based on the chemical rate equations that describe analogous processes in biological systems.¹¹

In the case of an adaptable learning system, the design approach must envisage not only the device, the assembled ensemble of devices, and all their associated variations, margins and susceptibilities, but it must also simulate the training process and tolerances associated with that stage. It will be important not to underestimate the challenge of this stage, although (compared to the situation when Si integrated circuits first appeared) the distributed computing power available to address the challenge is formidable!

Another major aspect of design approach challenges would arise if and when device/interconnect technologies presented the ability to create real 3-dimensional circuits. The types of design tools currently in use, and to a large extent, also the designers' thinking, are very much confined to slight variants above and below a single active plane. Notwithstanding all the associated issues of interconnect delays, power dissipation, *etc.*, the conceptual challenge of tackling a circuit hundreds of layers deep will require a fundamentally new design approach and sophisticated (intelligent) optimization tools.

8. Packaging and environmental concerns

Typically the packaging of a new device functionality is the last thing to be considered, and as such is often derivative, and non-optimized. Since we have been discussing some fundamentally different new approaches, it is likely that conventional packaging solutions will be less than ideal. In particular, some of the approaches (e.g. organic semiconductors) have inherent advantages (flexibility, thinness) that distinguish them from their rigid and breakable predecessors. These advantages should lead us to anticipate some radical new packaging approaches where, for example; the circuit could be conformally applied to an inside curved surface of the equipment casing, or even apparel!

The range of domestic (or medical) applications could also impose a new set of packaging demands, such as disposability, recyclability, high tolerance of humid or corrosive atmospheres. It is an important field for simulation and development work and, as for the case of design environments, would be worthy of early attention, given the magnitude of the problems to be solved.

At the same time, it is possible that new materials and packaging solutions will alleviate the challenge of ecological disposal. On the other hand, if certain applications were to achieve volumes of, say, daily newspaper usage, the manufacturing process should probably include the provision of reprocessing facilities.

9. Summary of powerful ideas to examine

In terms of device and interconnect technologies, four main areas are evident:

- The ability to work with devices offering less precise '1' and '0' states than we presently use, with potential to go to much smaller dimensions and grouped behaviors and characteristics.
- Devices that offer more "neural" types of functionality (multi-inputs with adjustable weightings; learnable or programmable response functions and arrays).
- Programmability of interconnect, possibly with regenerative characteristics, especially to deal with the two device scenarios envisaged above.
- Technology capability to offer real three-dimensional functionality and interconnects.

From a design focus viewpoint, two areas appear as dominant challenges:

- Architectural approaches to complex re-organizable systems, especially with *gedanken* devices, to help classify simple yet usable learning functions, and anticipate (or even guide) the device technologies.
- Design tools to help designers cope with the challenge of visualizing functionality and partitioning in three physical dimensions. Perhaps the design tools will draw upon current "virtual reality" three-dimensional game technology!

As for a more general vision of the technological future, I anticipate three major areas of discussion:

- The arrival of a "post-digital age"? An era where we have mastered powerful technologies that are better adapted to handling decision-making at arbitrary thresholds, using modest amounts of processing capability, and with learning/redundancy characteristics — essentially some biological analogs. The need for this type of technology will be driven by growing demand for "intelligent interface" functions to ease human access to progressively more automated technology in the domestic and professional environment. Implicitly, though, some functionality would still be best handled by classical digital circuits, calling for compatibility between the new and old media!
- In terms of planning for future capability, rather than anticipating that "the technology will be able to put out 1.3-meter wafers by the year 2020, so how can we use that much material?" we should first try to determine if there is a need (and therefore an economic driving force) for that scale of capacity.
- Visualize a future scenario, such as a particular device or technology capability, and then (without too much attention to how it would actually be achieved) simulate the advantages it might confer on existing technologies. This "what-if" analysis could offer some useful guidelines for prioritizing competing approaches for effort to be applied. It may well be that some

relatively unglamorous research advances could substantially transform an existing industry by resolving a present limitation.

10. Acknowledgments

It is an honor to acknowledge the inspiring discussions with my 5 colleagues from the Nortel Institute "think-tank" held in October 1997: Hamid Bolouri, Serge Luryi, Trey Smith, Horst Stormer and Jimmy Xu. In addition I would like to recognize pertinent inputs and feedback from Herb Goronkin and Albert Zylbersztein, as well as many useful exchanges with colleagues at Nortel.

References

1. Gordon Moore, "An update on Moore's Law", talk given at the Intel Developer Forum, September 30, 1997.
2. Paul Jay, "Growing up in the shadow of a silicon 'older brother' ", in: S. Luryi, J. M. Xu, and A. Zaslavsky, eds., *Future Trends in Microelectronics: Reflections on the Road to Nanotechnology*, NATO ASI series Vol. E323, Dordrecht: Kluwer, 1996, p. 71.
3. Hamid Bolouri (Univ. of Hertfordshire and Caltech); Paul Jay (Nortel); Serge Luryi (SUNY at Stony Brook); Trey Smith III (IBM, currently with Compaq Computer Corporation); Horst Stormer (Lucent Technologies) and Jimmy Xu (Univ. of Toronto).
4. *The National Technology Roadmap for Semiconductors, 1997 Edition* (Semiconductor Industry Association, San Jose, CA, 1997).
5. Don Monroe, "The end of scaling: disruption from below," in this book.
6. Herb Kroemer, "All that glitters isn't silicon," in: S. Luryi, J. M. Xu, and A. Zaslavsky, eds., *Future Trends in Microelectronics: Reflections on the Road to Nanotechnology*, NATO ASI series Vol. E323, Dordrecht: Kluwer, 1996, p. 1.
7. Simon Sze, "Evolution of non-volatile semiconductor memory: from floating-gate concept to single-electron RAM," in this book.
8. T. Ohshima and R. A. Kiehl, "Operation of bistable phase-locked single-electron tunneling logic elements," *J. Appl. Phys.* **80**, 912 (1996).
9. Herb Goronkin et al., "Progress in quantum functional devices to overcome barriers to ULSI scaling," *Proc. IEEE GaAs IC Symp.* **16**, 9 (1994).
10. Gilles Horowitz, "Organic transistors — present and future," in: S. Luryi, J. M. Xu, and A. Zaslavsky, eds., *Future Trends in Microelectronics: Reflections on the Road to Nanotechnology*, NATO ASI series Vol. E323, Dordrecht: Kluwer, 1996, p. 315.
11. A. Rust, H. Bolouri, R. Adams, and S. J. George, "Developmental evolution of an edge detecting retina," in: *Proc. Int. Conf. Neural Networks (ICANN'98)*, Vol. 2, Sweden, September 1998, pp. 561-566.

Process Technology for Sub-0.1 μm Silicon Devices

Junichi Murota, Takashi Matsuura, and Masao Sakuraba

Laboratory for Electronic Intelligent Systems, Research Institute of Electrical Communication, Tohoku University, Sendai 980-8577, Japan

1. Introduction

Super self-aligned processes i.e., selective surface reaction processes are extremely important for the fabrication of sub 0.1 μm Si devices as well as Si-based quantum-effect devices. Heterostructure growth processes of $\text{Si}_{1-x}\text{Ge}_x$ are attractive for the improvement of Si device performance. Since atomically flat surfaces and interfaces must be maintained, low-temperature processing is indispensable to suppress thermal degradation due to unexpected reactions and impurity diffusion. Low-temperature processing requires not only a clean surface but also an ultraclean reaction environment. Improvements in the quality of gases and equipment have enabled ultraclean low-temperature processing.¹⁻⁵

Our final goal is the development of atomic-order surface reaction processes and creation of new functional Si-based devices on the sub 0.1 μm length scale, as shown Fig. 1. Using surface reaction processes, we have made high performance MOSFETs⁶ with $\text{Si}_{0.5}\text{Ge}_{0.5}$ channels formed at 500 $^{\circ}\text{C}$ and also super-self-aligned shallow junction MOSFETs with a 0.1 μm gate length by utilizing *in-situ* impurity doped $\text{Si}_{1-x}\text{Ge}_x$ selective epitaxy on the source/drain regions by CVD at 550 $^{\circ}\text{C}$.⁷⁻¹⁰

In this paper, we describe low-temperature epitaxial growth of *in-situ* doped $\text{Si}_{1-x}\text{Ge}_x$ films on Si(100) by ultraclean LPCVD. Control of the deposition rate, the Ge fraction x and the doping concentration are discussed within the framework of

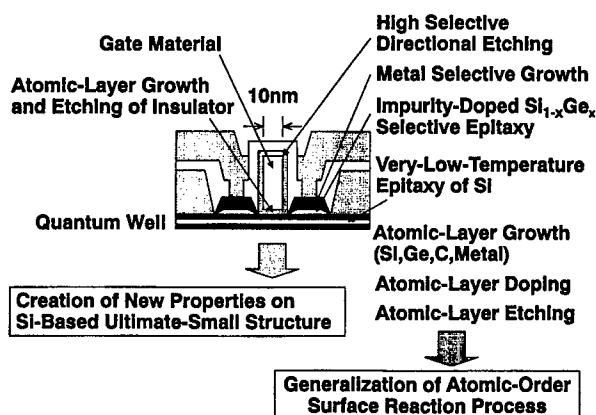


Figure 1. Si-based ultimate nanostructure fabrication.

the Langmuir-type adsorption and reaction scheme. Furthermore, atomic-order thermal nitridation of Si using NH_3 and self-limited atomic-layer growth of Si on the H-terminated Ge surface using SiH_4 are also explained by Langmuir-type kinetics. Finally, we propose atomic layer-by-layer deposition and etching with complete self-limiting separation of surface adsorption and reaction, as well as wide area activation using Xe flash heating and low energy ion irradiation.

2. Epitaxial growth of *in-situ* doped $\text{Si}_{1-x}\text{Ge}_x$ films by LPCVD

Low-temperature epitaxial growth of undoped and doped $\text{Si}_{1-x}\text{Ge}_x$ films on the Si(100) surface at 550 °C from SiH_4 , GeH_4 and H_2 with the B_2H_6 or PH_3 for dopant delivery was carried out in an ultraclean hot-wall LPCVD system.¹¹⁻¹³ The reaction rates of GeH_4 and SiH_4 are obtained from the deposition rates and Ge

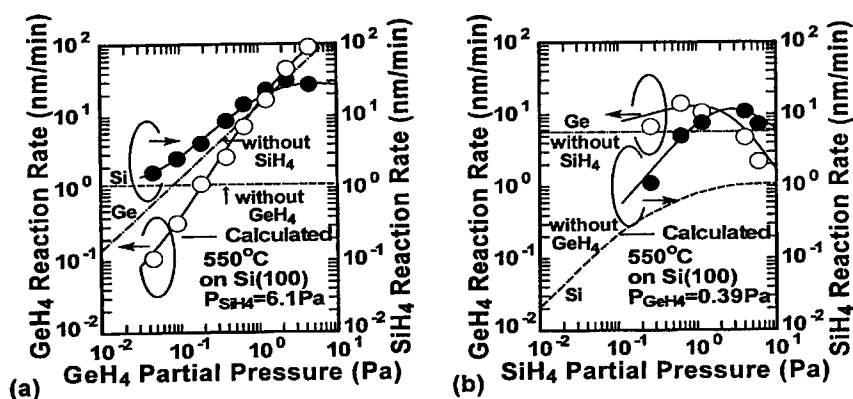


Figure 2. Dependences of the GeH_4 and SiH_4 reaction rates on the (a) GeH_4 and (b) SiH_4 partial pressures. The lines are calculated results.

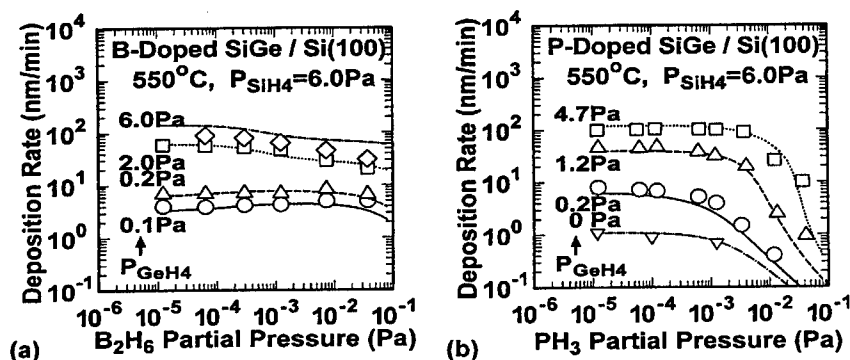


Figure 3. Dependences of the deposition rate on the (a) B_2H_6 and (b) PH_3 partial pressures. The lines are calculated results.

fraction in undoped $\text{Si}_{1-x}\text{Ge}_x$. With increasing GeH_4 partial pressure, the GeH_4 reaction rate increases monotonically, while the SiH_4 reaction rate increases up to the maximum value and then decreases, see Fig. 2(a). With increasing SiH_4 partial pressure, the GeH_4 and SiH_4 reaction rates increase up to the maximum value and then decrease, see Fig. 2(b). The SiH_4 and GeH_4 reaction rates are expressed by a Langmuir-type rate equation, assuming that one SiH_4 or GeH_4 molecule is adsorbed at a single adsorption site and decomposes there. We find that the SiH_4 and GeH_4 adsorption rate constants are largest at the bond site of the Si-Ge pair, while the SiH_4 surface reaction rate constant becomes the largest at the bond site of the Ge-Ge pair.

In the case of B doping, the reduction of the deposition rate occurs at the higher GeH_4 partial pressure, see Fig. 3(a). The Ge fraction scarcely changes with the B_2H_6 addition (Fig. 4(a)). The B concentration is nearly proportional to the B_2H_6 partial pressures (Fig. 5(a)). In the case of P doping, the reduction of the deposition rate shifts to the higher PH_3 partial pressures with increasing GeH_4 partial pressure (Fig. 3(b)). The Ge fraction x increases in the higher PH_3 partial

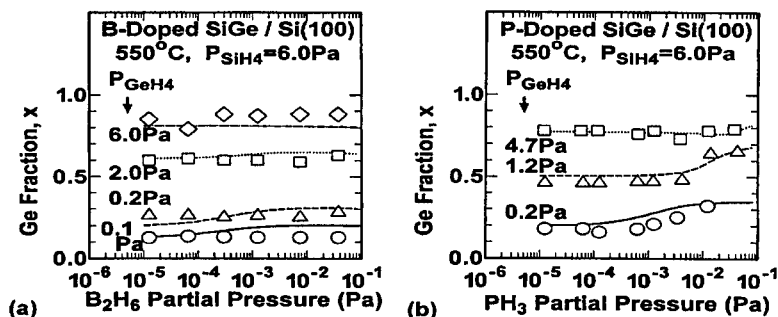


Figure 4. Dependences of the Ge fraction on the (a) B_2H_6 and (b) PH_3 partial pressures. The lines are calculated results.

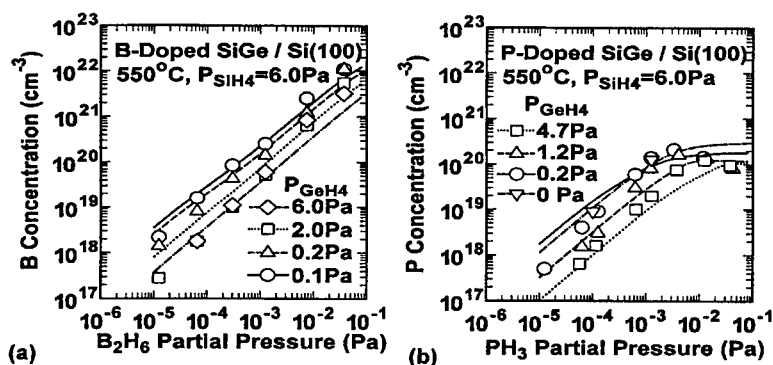


Figure 5. Dependences of the dopant concentration on the (a) B_2H_6 and (b) PH_3 partial pressures. The lines are calculated results.

$$R_{Si} = \frac{k_1 P_{SiH_4} n_0}{1 + (k_1 / k_{Si}) P_{SiH_4}} \quad (1)$$

$$R_{Ge} = \frac{k_2 P_{GeH_4} n_0}{1 + (k_1 / k_{Si}) P_{SiH_4}} \quad (2)$$

$$x = \frac{R_{Ge}}{R_{Si} + R_{Ge}} \quad (3)$$

$$k_a n_0 = \sum_{i=1}^3 k_{ai} n_0 c_i, \quad (a = 1, 2, Si) \quad (4)$$

$$c_1 = (1-x)^2, \quad c_2 = 2x(1-x), \quad c_3 = x^2 \quad (5)$$

$$k_a n_0 = \sum_{i=1}^3 k_{ai} n_0 c_i \frac{1}{1 + K_{Di} P_D}, \quad (a = 1, 2, Si) \quad (6)$$

$$K_{Di} = \frac{k_{Di}}{K_{SDi} (R_{Si} + R_{Ge}) c_i + k_{-Di}}, \quad (D = P, B) \quad (7)$$

$$C_D = \sum_{i=1}^3 K_{SDi} c_i \frac{K_{Di} P_D}{1 + K_{Di} P_D}, \quad (D = P, B) \quad (8)$$

Table 1. Equations in the Langmuir-type adsorption and reaction scheme for formulating *in-situ* doping in $Si_{1-x}Ge_x$.

pressure region (Fig. 4(b)). The P concentration increases up to a maximum value and then decreases with increasing PH_3 partial pressure, see Fig. 5(b). Referring to the Langmuir-type adsorption and reaction scheme, these doping characteristics can be explained if it is assumed that one dopant molecule (B-hydride or P-hydride) occupies one free surface site according to Langmuir's adsorption isotherm, that the site where the dopant molecule is adsorbed becomes inactive for both the SiH_4 and GeH_4 adsorption/reactions on the surface, that such dopant occupancy is different at the Si-Si, Si-Ge and Ge-Ge pair sites, and that the dopant incorporation in the grown film obeys Henry's law. Moreover, at higher PH_3 partial pressures, lower solid solubility of P on the Si-Ge and Ge-Ge pair sites than that on the Si-Si pair site may be the additional origin of the decrease of the P concentration.

3. Atomic-order surface reaction on H-terminated Si and Ge surfaces

A wet-cleaned Si(100) surface formed by diluted HF solution dipping is terminated by hydrogen.¹⁴ We found that preheating under various conditions produces dihydride, monohydride, reconstructed mono-hydride with dimer structure, and H-

free surfaces of Si(100) and Ge(100).^{15,16} Experiments on W growth from a WF_6 - SiH_4 gas mixture^{17,18} and Ge epitaxial growth from GeH_4 gas¹⁹ indicate that the incubation period in film growth on Si at a very low temperature arises from reaction suppression due to surface H-termination. It is also observed that CH_4 reacts on the H-free surface forming a self-limited atomic-layer of C at 600 °C or lower according to the Langmuir-type kinetics.²⁰ In this section, we discuss atomic-order surface reactions on the H-terminated surface.

- *Low-temperature nitridation of Si using NH_3* ^{21,22}

The NH_3 exposure time dependence of the N atom concentration on Si(100) for different pretreatments is compared in Fig. 6. On the Si surface after wet-cleaning, FTIR/RAS spectra showed that the Si-hydride coverage decreases with increasing NH_3 exposure time and becomes below the detection limit just when the N atom concentration becomes the surface monolayer ($\sim 6 \times 10^{14} \text{ cm}^{-2}$). These results indicate that the NH_3 reaction on Si(100) proceeds with desorption of the hydrogen atom bonded to the top Si atom and that the N atoms cover the Si surface instead of the hydrogen atoms. In this case, it is believed that NH_3 molecules were adsorbed on the Si surface but did not react. Now, assuming that NH_3 is adsorbed physically at a single adsorption site according to the Langmuir-type adsorption isotherm, and that the NH_3 surface reaction proceeds according to a first-order reaction including an initial small and very fast reaction, the N atom concentration can be expressed as a function of the NH_3 pressure in very good agreement with the experimental data, as shown in Fig. 6.

After preheating in an Ar environment at 650 °C, the N atom concentration on the Si surface increased spontaneously up to $\sim 2 \times 10^{14} \text{ cm}^{-2}$. It is clear that the H-free Si surface is formed by Ar-preheating at 650 °C. FTIR/RAS spectra showed that the Si-monohydride is generated and increases with increasing NH_3 exposure time up to 10 min and then decreases. Therefore, we believe that NH_3 is dissociated into NH_x ($x = 0, 1, 2$) and H atoms are adsorbed initially on the Si dangling bonds. Further nitridation may proceed with H desorption.

- *Atomic-layer growth of Si on Ge(100) using SiH_4* ²³⁻²⁵

A self-limited atomic layer of Si was deposited on the Ge surface at 260 °C at the SiH_4 pressure of 500 Pa. The time and pressure dependence of the adsorption and reaction of SiH_4 can be explained by Langmuir-type kinetics, as shown in Fig. 7. The saturated surface concentration of the deposited Si atoms depends on the preheating conditions. Single atomic layer growth of Si was seen for the H-free surface formed by preheating at 350 °C in Ar, and for the H-terminated unreconstructed surface by preheating at 260 °C in H_2 . However, in the case of preheating in H_2 at 350 °C, where an H-terminated dimer structure forms on the Ge surface, the Si atom concentration hardly reached that of a single atomic layer. Thus, we conclude that the density of the SiH_4 reaction sites on the H-terminated surface with the dimer structure is lower than that of the H-terminated unreconstructed surface or the H-free surface.

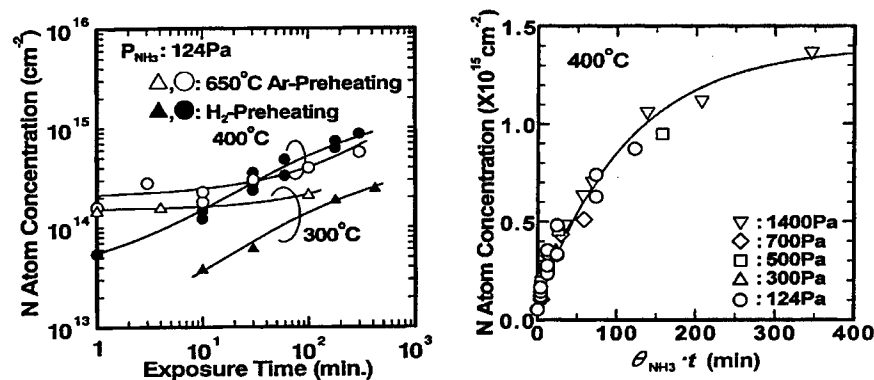


Figure 6. NH_3 exposure time dependence of the N atom concentration on Si(100) (left) and θ_{NH_3} -time product dependence of the N atom concentration on the wet-cleaned Si(100) (right). The solid line is a fit by $n_{\text{N}} = (N_{\text{sat}} - N_i)[1 - \exp(-k_2 \theta_{\text{NH}_3} t)] + N_i$, where $\theta_{\text{NH}_3} = k_1 P_{\text{NH}_3} / (k_1 P_{\text{NH}_3} + k_{-1})$.

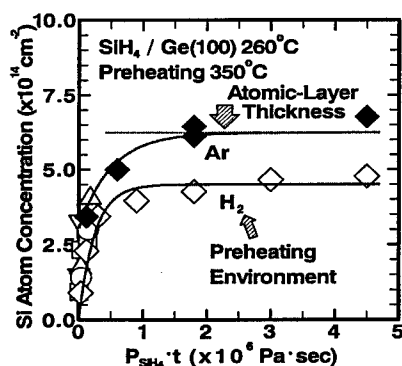


Figure 7. Pressure-time product dependence vs. Si atom concentration on Ge(100) after SiH_4 exposure. Data points correspond to SiH_4 partial pressures: 10 Pa (inverted triangles), 29 Pa (triangles); 100 Pa (circles); 300 Pa (open diamonds); 500 Pa (filled diamonds). Solid lines are fits by $N_{\text{S}} = (N_{\text{sat}} - N_i)[1 - \exp(-k_{\text{SiH}_4} P_{\text{SiH}_4} t)]$.

4. Atomic layer-by-layer processing of Si-Ge

Atomic layer-by-layer processing of a Si-Ge system is important for fabricating future semiconductor devices, e.g. single or multiple quantum well or superlattice structure devices. In a conventional surface reaction process, adsorption and reaction proceed simultaneously. In order to control a process with atomic-layer precision, it is important to separate the adsorption and the reaction. A self-limiting mechanism is essential, because it necessarily gives a constant and stable

process determined by the thermodynamic saturation condition. In this section, we propose atomic layer-by-layer epitaxy and etching with complete self-limiting separation of surface adsorption and reaction, as well as wide area activation at once using Xe flash heating and low energy ion irradiation, and also discuss their adsorption mechanism using Langmuir-type kinetics.

• *Epitaxy by CVD*^{11,23,24,26,27}

The separation between surface adsorption and reaction of SiH_4 or GeH_4 on a Si substrate has been achieved by flash-heating the surface with a Xe flash lamp using an ultraclean rf-heated, cold-wall low-pressure CVD system. While rf-heating the substrate, SiH_4 or GeH_4 gas is introduced into the reactor, and then the reactant molecules adsorbed on the surface are decomposed by flash-lamp light shots (1 ms, 20-60 J/cm²). Since a Si substrate loaded after wet cleaning was not heated above 400 °C, the Si substrate surface was still H-terminated.

The thickness deposited per shot became independent of the substrate temperature in the range of 385-395 °C for Si layer growth and of 260-275 °C for Ge layer growth as shown in Fig. 8. In these temperature ranges, the shot-to-shot time interval dependence of the film thickness deposited per shot using SiH_4 and GeH_4 is shown in Fig. 9. It is found that the deposited thickness increases and then saturates with the time interval. This saturation means that continuous SiH_4 or GeH_4 decomposition during the interval is negligible, implying self-limited SiH_4 or GeH_4 adsorption. As shown in Fig. 9(a), the deposited Si thickness in the saturation region depends on the SiH_4 partial pressure. This means that the saturated amount of adsorbed SiH_4 molecules is determined by the balance between adsorption and desorption. On the other hand, the deposited Ge thickness in the saturation region equals the single-atomic-layer thickness — see Fig. 9(b).

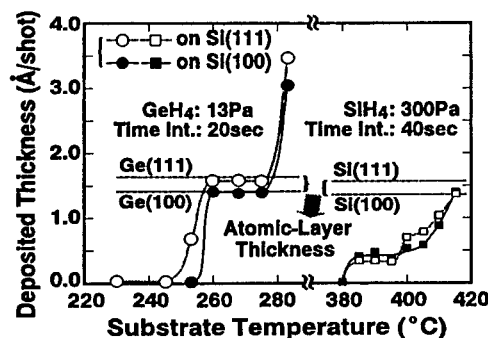


Figure 8. Substrate temperature dependence of the film thickness deposited per flash-lamp light shot.

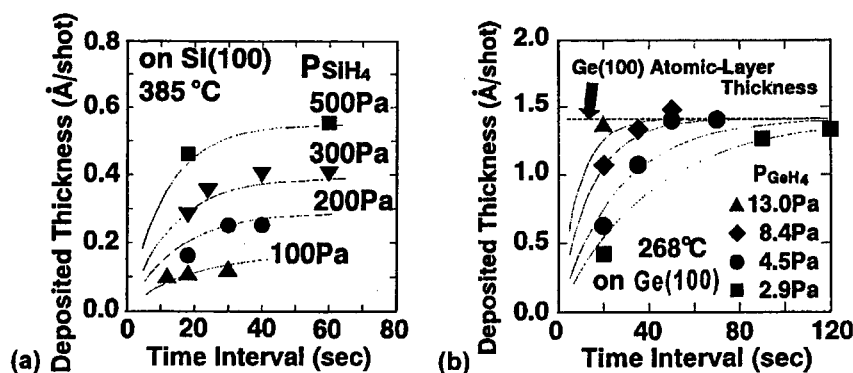


Figure 9. Shot-to-shot time interval dependence of the film thickness deposited per flash-lamp light shot using (a) SiH₄ and (b) GeH₄. The solid lines are fits by $(\text{deposited thickness}) = (\text{atomic layer thickness}) (n_0/N) [k_{\text{MH}_4} P_{\text{MH}_4} / (k_{\text{MH}_4} P_{\text{MH}_4} + k_{-\text{MH}_4})] \times [1 - \exp(-(k_{\text{MH}_4} P_{\text{MH}_4} + k_{-\text{MH}_4})t)]$, where (n_0/N) is the total adsorption site density normalized by the surface atom density and the $k_{-\text{MH}_4}$ desorption term is included for Si but negligible for Ge.

Calculations based on Langmuir's kinetics are in excellent agreement with the experimental data, suggesting that the total adsorption site density equals the surface atom density and a SiH₄ or GeH₄ molecule occupies only one adsorption site. The single-atomic-layer growths per shot of Si and Ge are expected in the SiH₄ and GeH₄ partial pressure ranges of above a few thousand Pa and a few Pa, respectively.

A Si film was deposited on the Ge(100) surface without flash shot even at temperatures below 300°C, as described above. Conversely, the incubation period of Ge growth was a few tens of shots. A Ge layer was scarcely deposited by the first single-shot on the wet-cleaned Si surface at the GeH₄ partial pressure of 23 Pa, but completely at 300 Pa. Therefore, we believe that the GeH₄ adsorption rate on the Si surface is lower than on the Ge surface. These results show that single-atomic-layer epitaxy of Si and Ge is possible with CVD.

- *Etching by plasma*²⁸⁻³¹

Chlorine molecules and/or radicals were used as adsorbants for Si etching, while only chlorine molecules for Ge etching. Reaction was induced by irradiation of low energy Ar⁺ ions generated by an ultraclean ECR plasma apparatus. Atomic layer-by-layer etching was achieved by cyclically repeating these chlorine adsorption and ion induced reactions. Almost no undercut below the mask was observed, which demonstrates the merit of utilizing ion-induced reactions.

The atomic-layer etch rate per cycle for Si is controlled by the product of the chlorine supply time and partial pressure (i.e., the chlorine exposure). The rate increases with exposure, and tends to saturate at a fractional atomic-layer

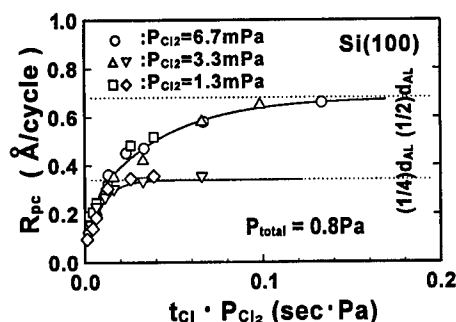


Figure 10. Si etch rate per cycle vs. chlorine exposure. The solid lines are fits by $R_{pc} = R_{pcsat} [1 - \exp(-kP_{Cl_2}t_{Cl})]$.

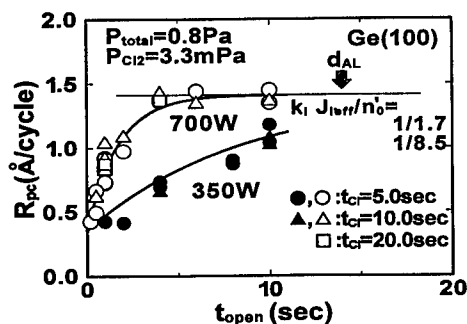


Figure 11. Dependence of the Ge etch rate per cycle on Ar^+ ion irradiation time. Solid lines are fits using $R_{pc} = d_{AL} [1 - (3/4)\exp(-(k_i J_{ion}/n_0)t_{open})]$.

thickness, as shown in Fig. 10. The saturation value in the case with chlorine radicals is twice as large as that with chlorine molecules only. The saturation value also depends on substrate orientation, where the correspondence was observed with the ratio of the number of the available chlorine adsorption sites per one Si atom to the average number of the remaining Si-Si bonds per Si atom.

The atomic-layer etch rate per cycle for Ge increases with the Ar^+ ion irradiation time and tends to saturate to a single atomic layer thickness, as shown in Fig. 11, while chlorine adsorption on Ge is too fast to observe chlorine exposure dependence in the present conditions. The saturation characteristics for Ge are derived by assuming the simplest Ar^+ -induced reaction with a single effective reaction probability. Moreover, by comparing the time necessary for saturation at different plasma exciting powers to the ion flux density measurements, we estimate that ~ 10 Ar^+ ions with an energy higher than ~ 13 eV are required to remove a Ge atom with chlorine adsorption. From these results, it is proposed that single-atomic layer-by-layer etching is possible with plasma processing.

5. Acknowledgments

The authors acknowledge support by a Grant-in-Aid for Scientific Research from the Ministry of Education, Science, Sports and Culture of Japan, Japan Society for Promotion of Science (JSPS-RFTF96R13101) and the Mitsubishi Foundation.

References

1. J. Murota, N. Nakamura, K. Kato, *et al.*, "Low-temperature silicon selective deposition and epitaxy on Si using thermal decomposition of silane under ultraclean environment," *Appl. Phys. Lett.* **54**, 1007 (1989).
2. T. Matsuura, H. Uetake, T. Ohmi, *et al.*, "Directional etching of Si with perfect selectivity to SiO₂ using an ultraclean electron cyclotron resonance plasma," *Appl. Phys. Lett.* **56**, 1339 (1990).
3. H. Uetake, T. Matsuura, T. Ohmi, *et al.*, "Anisotropic etching of n^+ polysilicon with high selectivity using a chlorine and nitrogen plasma in an ultraclean ECR etcher," *Appl. Phys. Lett.* **57**, 596 (1990).
4. J. Murota, M. Kato, R. Kircher, and S. Ono, "Low-temperature Si and Ge CVD in ultraclean environment," *J. Physique IV* **1**, C2-795 (1991).
5. T. Matsuura, T. Ohmi, J. Murota, and S. Ono, "Inversion from selective homoepitaxy of Si to selective Si film deposition on SiO₂ using an ultraclean electron cyclotron resonance plasma," *Appl. Phys. Lett.* **61**, 2908 (1992).
6. K. Goto, J. Murota, T. Maeda, *et al.*, "Fabrication of a Si_{1-x}Ge_x channel MOSFET containing high Ge fraction layer by low-pressure chemical vapor deposition," *Jpn. J. Appl. Phys.* **32**, 438 (1993).
7. J. Murota, M. Kato, N. Mikoshiba, and S. Ono, "Selective epitaxial growth of Si and Ge at low temperatures using ultraclean CVD processing," *Int. Conf. SSDM*, Yokohama, Japan (1991), p. 141.
8. F. Honma, J. Murota, K. Goto, T. Maeda, and Y. Sawada, "Ultrashallow junction formation using low-Temperature selective Si_{1-x}Ge_x chemical vapor deposition," *Jpn. J. Appl. Phys.* **33**, 2300 (1994).
9. K. Goto, J. Murota, F. Honma, *et al.*, "A novel fabrication method for short channel MOSFETs using self-aligned ultrashallow junction formation by selective Si_{1-x}Ge_x CVD," *Int. Conf. SSDM*, Yokohama, Japan (1994), p. 999.
10. J. Murota, M. Ishii, K. Goto, *et al.*, "Fabrication of 0.1 μ m MOSFET with super self-aligned ultrashallow junction electrodes using selective Si_{1-x}Ge_x CVD," *Proc. 27th Europ. Solid-State Dev. Res. Conf.*, Stuttgart (1997), p. 376.
11. J. Murota and S. Ono, "Low-temperature epitaxial growth of Si/Si_{1-x}Ge_x/Si heterostructure by CVD," *Jpn. J. Appl. Phys.* **33**, 2290 (1994).
12. J. Murota, Y. Takasawa, H. Fujimoto, *et al.*, "Low-temperature epitaxial growth mechanism of Si_{1-x}Ge_x films in the silane and germanium reactions," *J. Physique IV* **5**, C5-1165 (1995).
13. J. Murota, A. Moriya, M. Sakuraba, *et al.*, "In-situ heavy doping of P and B in low-temperature Si_{1-x}Ge_x epitaxial growth using ultraclean LPCVD," *Proc.*

- 8th Int. Symp. Silicon Mater. Sci. Technol., San Diego (1998), p. 822.
14. Y. J. Chabal, "Infrared study of the chemisorption of hydrogen and water on vicinal Si(100) 2 x 1 surfaces," *J. Vac. Sci. Technol. A* **3**, 1448 (1985).
 15. M. Sakuraba, J. Murota, and S. Ono, "Stability of the dimer structure formed on Si(100) by ultraclean low-pressure CVD," *J. Appl. Phys.* **75**, 3701 (1994).
 16. M. Sakuraba, T. Matsuura, and J. Murota, "H-termination on Ge(100) and Si(100) by diluted HF dipping and by annealing in H₂," *Proc. 5th Int. Symp. Cleaning Technol. Semicond. Dev. Manufact.*, Paris (1997), p. 213.
 17. Y. Yamamoto, T. Matsuura, and J. Murota, "Selective growth of W at very low temperatures using a WF₆-SiH₄ gas system," *Proc. 13th Int. CVD Conf.* Los Angeles (1996), p. 814.
 18. Y. Yamamoto, T. Matsuura, and J. Murota, "Surface reaction of alternately supplied WF₆ and SiH₄ gases," *Surf. Sci.* **408**, 190 (1998).
 19. S. Kobayashi, M. Sakuraba, T. Matsuura, *et al.*, "Initial growth characteristics of Ge on Si in LPCVD using germane," *J. Crystal Growth* **174**, 686 (1997).
 20. A. Izena, M. Sakuraba, T. Matsuura, and J. Murota, "Low-temperature reaction of CH₄ on Si(100)," *J. Crystal Growth* **188**, 131 (1998).
 21. T. Watanabe, M. Sakuraba, T. Matsuura, and J. Murota, "Atomic-order layer growth of silicon nitride films at low temperatures," *Proc. 13th Int. CVD Conf.*, Los Angeles (1996), p. 504.
 22. T. Watanabe, M. Sakuraba, T. Matsuura, and J. Murota, "Atomic-order nitridation of the H-terminated and H-free Si surfaces by NH₃," *Proc. 14th Int. CVD Conf.*, Paris (1997), p. 97.
 23. M. Sakuraba, J. Murota, T. Watanabe, Y. Sawada, and S. Ono, "Atomic-layer epitaxy control of Ge and Si in flash-heating CVD using GeH₄ and SiH₄ gases," *Appl. Surf. Sci.* **82**, 354 (1994).
 24. J. Murota, M. Sakuraba, T. Watanabe, *et al.*, "Atomic layer-by-layer epitaxy of Si and Ge using flash heating in CVD," *J. Physique IV* **5**, C5-1101 (1995).
 25. T. Watanabe, M. Sakuraba, T. Matsuura, and J. Murota, "Atomic-layer surface reaction of SiH₄ on Ge(100)," *Jpn. J. Appl. Phys.* **36**, 4042 (1997).
 26. M. Sakuraba, J. Murota, N. Mikoshiba, and S. Ono, "Atomic layer epitaxy of Ge on Si using flash heating CVD," *J. Crystal Growth* **115**, 79 (1991).
 27. J. Murota, M. Sakuraba, and S. Ono, "Silicon atomic layer growth controlled by flash heating in CVD using SiH₄ gas," *Appl. Phys. Lett.* **62**, 2353 (1993).
 28. T. Matsuura, J. Murota, Y. Sawada, and T. Ohmi, "Self-limited layer-by-layer etching of Si by alternated chlorine adsorption and Ar⁺ ion irradiation," *Appl. Phys. Lett.* **63**, 2803 (1993).
 29. K. Suzue, T. Matsuura, J. Murota, Y. Sawada, and T. Ohmi, "Substrate orientation dependence of self-limited atomic-layer etching of Si with chlorine adsorption and low-energy Ar⁺ irradiation," *Appl. Surf. Sci.* **82-83**, 422 (1994).
 30. T. Sugiyama, T. Matsuura, and J. Murota, "Atomic-layer etching of Ge using an ultraclean ECR plasma," *Appl. Surf. Sci.* **112**, 187 (1997).
 31. T. Matsuura, T. Sugiyama, and J. Murota, "Atomic-layer surface reaction of chlorine on Si and Ge assisted by an ultraclean ECR plasma," *Surf. Sci.* **402-404**, 202 (1998).

Hot Carrier Degradation Issues in Submicron MOSFETs

K. Hess, L. F. Register, B. Tuttle, J. Lyding

*Beckman Institute, Coordinated Science Laboratory, University of Illinois at
Urbana-Champaign, Urbana, IL 61801 U.S.A.*

I. C. Kizilyalli

Lucent Bell Laboratories, Murray Hill, NJ 07974 U.S.A.

1. Introduction

An important hot-carrier effect in semiconductor devices today is hot-carrier degradation, which limits the operating lifetime of MOS (metal-oxide-silicon) devices. For more than twenty-five years there has been ongoing research in this field, yet such degradation remains a serious and persistent limitation on MOS transistor design.¹ During that time frame, however, the understanding of hot-carrier degradation has progressed far less than the understanding of carrier transport. Even the basic mechanisms of degradation remain subjects of controversy, and modeling of degradation remains an essentially empirical process. However, reliable degradation modeling is arguably more important than transport modeling to the development of future generations of submicron MOS technologies, at least in one critical respect. Device characteristics such as threshold (switch-on) voltage, frequency response, off-state leakage currents, etc. are measured to confirm or contradict the predictions of models before products go to market, but there is no direct way to measure device lifetimes that under normal operating conditions exceed product development times; the semiconductor industry must rely on predictions of device degradation.

As a semi-experimental alternative to direct measurement of device lifetimes under normal operating conditions, it has become standard practice to measure device lifetimes under varying accelerated stress conditions — that amount to applying higher than normal operating voltages — and then extrapolate the results back to normal operating conditions. However, not only are the extrapolations made without aid of a strong first-principles understanding of the degradation mechanism, but it has been shown that the degradation mechanism itself can depend on the degree of accelerated stressing.² Therefore, in principle, serious discrepancies between predicted and actual device lifetimes could occur and go undiscovered through a number of product generations, and with medical, aeronautical, military and other critical applications one could only hope that the method of discovery would be relatively benign. This is not to say that accelerated stress testing is not an effective tool, quite to the contrary, but to point out that

predicting device reliability is a risky endeavor and that there is no guarantee that this approach will always work as well for future generations of submicron devices.

While no approach could eliminate the risk associated with predicting device lifetimes completely, an improved first-principles understanding of device degradation and the addition of associated simulation tools could help to minimize it, and also be of considerable aid during design stages. In this article, we first review recent advances in the first-principles understanding of at least one form of degradation, surface depassivation. These advances have been achieved by going far beyond conventional approaches into atomistic properties. Then we identify the remaining challenges to the development of a first-principles-based modeling capability for degradation of MOS devices, and present preliminary efforts to address these challenges.

2. Depassivation and the hydrogen/deuterium isotope effect

MOS devices are processed with hydrogen (H) to saturate the always present dangling bonds at the silicon/silicon-dioxide (Si/SiO₂) interface to improve and reduce variation in device performance. However, this passivation also sets the stage for subsequent degradation. Although there may be multiple contributions to degradation in MOS devices, an important and often dominant component had been widely believed, if not actually confirmed, to be hot-carrier induced depassivation of the (Si/SiO₂) interface over time.^{3,4} The associated charge trapping at the interface would explain shifts in the threshold voltage and degradation of the channel conductance over time, and, with the continuing reductions in supply voltages and the ratio of conduction channel volume to surface area, would suggest that future generations of submicron MOS devices will only become more sensitive to such surface degradation.

Although surface depassivation was widely believed to be an important source of hot-carrier degradation, the mechanism of this depassivation had remained a subject of controversy. However, recent scanning tunneling microscopy (STM) experiments on passivated silicon surfaces in a vacuum have shed new light on this subject, and demonstrated a giant isotope effect between hydrogen and deuterium (D) passivated surfaces. Indeed, these experiments have identified two distinct mechanisms by which depassivation can occur,⁵ as depicted in Fig. 1, each with its own isotope effect favoring reduced desorption of D.⁶ At high voltages (injected electron energies) and low currents, apparently excitation of the Si-H/D bond electron to an excited state is responsible for a force that repels and desorbs the H or D atom. This process is similar to the well-known electronic surface desorption or photochemistry, as described by Menzel and Gomer⁷ among others. The observed 50-fold isotope effect is likely due to the greater difficulty of accelerating the heavier D away from the surface during the lifetime of the excited state of the bond electron. At lower voltages and higher currents, bond-breaking is probably via excitation of the vibrational mode(s) of the Si-H/D bond (bond

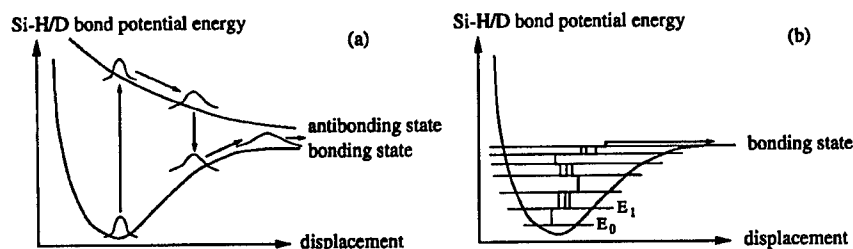


Figure 1. Two physical mechanisms of desorption. (a) Excitation of the bond electron from a bonding to an antibonding state accelerates the H/D atom away from the bond, potentially providing enough energy during the excited state lifetime to allow desorption. (b) Multiple excitations of the vibrational modes of the bonding state "heats" the bond increasing the probability of thermal desorption.

heating) caused by multiple collisions with the STM electrons within the lifetime of the vibrational mode(s); in contrast to the previous mechanism, which has an almost current-independent yield, this mechanism exhibits a strong current dependence characteristic of a process requiring multiple scattering events. In this latter case, the still larger observed isotope effect is probably due to the shorter lifetime of at least the bending vibrational mode for D which is more closely matched to the vibrational modes of bulk silicon.⁸

Motivated by the observed isotope effect in the STM experiments and the similarities between the STM and MOS systems, an analogous experiment was performed for MOS field-effect transistors (MOSFETs); n-channel MOSFETs were processed both with H and D and then accelerated stress tests were performed.⁹ A typical result for transistors of CMOS technology is shown in Fig. 2. The large isotope effect and its technical potential are obvious and have since been verified extensively. This isotope effect also represents perhaps the best experimental confirmation of the role of surface depassivation in the aging of MOS devices.

However, there are nontrivial differences between the STM and MOS systems that must be examined before a comprehensive first-principles understanding of, and an associated modeling capability for, hot-carrier-induced surface depassivation in MOS devices will be possible. The monoenergetic carriers of the STM system are replaced by complicated hot-carrier distributions in the conduction channel of MOSFETs that are sensitive to both device geometry and bias conditions, and the regular array of dangling silicon bonds in a vacuum environment are replaced by isolated dangling bonds in the amorphous oxide environment. Thus, to address depassivation in MOS devices will require merging

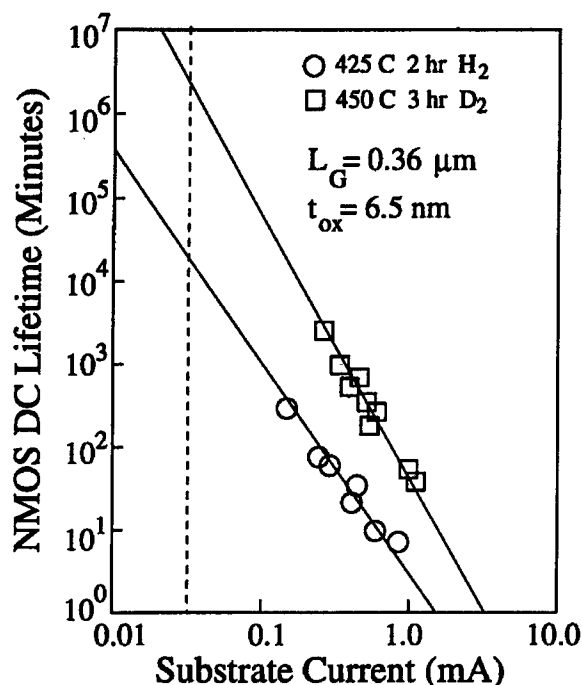


Figure 2. Hot-electron degradation lifetime for 10% shift in the threshold voltage as a function of substrate current. The dashed line indicates normal operating conditions. Extrapolation from the accelerated stress tests suggests a lifetime of less than a month for the H-passivated device, but approximately 4 years for the D-passivated device. The measurements were taken on a fully processed wafer with several layers of metallization.

a number of diverse fields including hot-carrier transport, molecular bonding and dynamics, and scattering theory as required to model the coupling between hot carriers and the Si-H/D bond.

3. Channel hot-carrier-induced desorption

To explore the effects of the channel hot-carrier source but not yet those of the oxide environment on the bond itself, a simple model of depassivation was developed and calibrated to the STM data and then applied to the MOS system.

As a first approximation the two desorption processes of Fig. 1 are treated as separable such that the total desorption rate is simply,

$$D \cong D_e + D_v, \quad (1)$$

where D_e is the desorption rate via excitation of the bonding electron from the bonding to an antibonding state, and D_v is desorption via heating of the vibrational modes of the bonding state. Even if there is mixing of these processes, Eq. (1) still establishes a lower boundary on the desorption rate.

For desorption via excitation of the bonding electron, the desorption rate D_e can be written,

$$D_e \simeq \int_{E_{\text{thres}}}^{\infty} dE I(E) \bar{\sigma}(E) P(E), \quad (2)$$

where $I(E)$ is the carrier impact frequency on the surface per unit area per unit energy, $\bar{\sigma}(E)$ is the scattering cross-section for excitation of the bound electron (with a weighted average over the angle of incidence), $P(E)$ is the probability that excitation of the bond's electronic state will actually lead to desorption, and E_{thres} is the threshold energy for scattering.

For desorption via excitation of the vibrational mode(s) of the bond, i.e., bond heating, the cumulative effect of multiple carrier impacts must be addressed. A model that has proven successful in explaining the STM data⁵ is the truncated harmonic oscillator model for which the desorption rate is given by,

$$D_v \simeq \left\{ \left(\frac{E_b}{\hbar\omega} + 1 \right) [R_{\text{em}} + \exp(-\hbar\omega/k_B T_L)/\tau] \right\} \times \left[\frac{R_{\text{ab}} + 1/\tau}{R_{\text{em}} + \exp(-\hbar\omega/k_B T_L)/\tau} \right]^{-(E_b/\hbar\omega)}, \quad (3)$$

where $\hbar\omega$ is the phonon energy of the Si-H/D bond, E_b is the barrier height to desorption, T_L is the background lattice temperature, and τ is the phonon lifetime. R_{em} is the total phonon emission rate divided by the phonon occupation number plus one,

$$R_{\text{em}} \simeq \int_{\hbar\omega}^{\infty} dE I(E) \bar{\sigma}_{\text{em}}(E) [1 - f(E - \hbar\omega)], \quad (4)$$

and R_{ab} the total phonon absorption rate divided by the phonon occupation number,

$$R_{\text{ab}} \simeq \int_0^{\infty} dE I(E) \bar{\sigma}_{\text{ab}}(E) [1 - f(E + \hbar\omega)], \quad (5)$$

where $\bar{\sigma}_{\text{em}}(E)$ and $\bar{\sigma}_{\text{ab}}(E)$ are the scattering cross-sections for bond-phonon emission and absorption, respectively. While the truncated harmonic oscillator model could be considered oversimplified, particularly near the barrier top, the model produces a quite reasonable result: the desorption rate depends exponentially on a bond temperature established by the competition between coupling to "hot" carriers and the relatively cool crystal lattice.

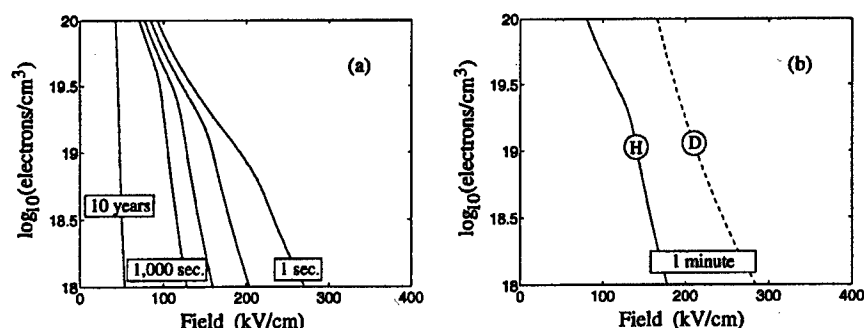


Figure 3. Contours of 10% depassivation as a function of electric field and interface carrier density for (a) H depassivation as a function of lifetime and (b) H and D depassivation for the same lifetime. Regions of field and carrier density to the upper right of the contours for a given amount of time are associated with greater than 10% depassivation; regions to the lower left, less depassivation. In (a) upper sections of the contour lines for 1–1,000 seconds (accelerated stress conditions) are associated with bond heating. In (b) note the significant isotope effect and the lack of bond-heating induced depassivation for deuterium.

To isolate the effects of the channel hot-carrier source — and because there is little option at this point — the scattering cross-sections and other model parameters are obtained by fitting to the STM data⁶ (as described in somewhat greater detail in Ref. 10) ignoring the differences in the Si-H/D bonds between the STM and MOS systems. What remains to be addressed are the differing carrier impact frequencies on the surface per unit area per unit energy $I(E)$ between the STM and MOS systems. For this work $I(E)$ is obtained as a function of carrier density and field within ranges representative of the conduction channel of MOSFETs using published results of semiclassical Monte Carlo simulations.¹¹ While in submicron MOS a model of the carrier energy distribution that depends only on the field is an oversimplification, this approximation provides a good starting point for exploring the qualitative effects of interest here. In practice, calculation of the surface impact rate as a function of device geometry and applied voltage under normal or accelerated stress conditions, along with consideration of feedback of the surface degradation on this surface impact rate, will be required.

Typical results of this analysis are shown in Fig. 3. The results of Fig. 3(a) suggest that both desorption mechanisms identified in the STM experiments are plausible sources of surface degradation in MOS devices as well. Figure 3(a) also exhibits the bond-heating mechanism to be more pronounced at higher fields and carrier densities, suggesting caution in the extrapolation of device lifetimes under normal operating conditions (years presumably) from those obtained under accelerated stress testing (minutes). Figure 3(b) displays the isotope effect, indicating, consistent with experiments,⁹ the possibility of much longer MOSFET lifetimes using D under the same applied biases, or the possibility of significantly

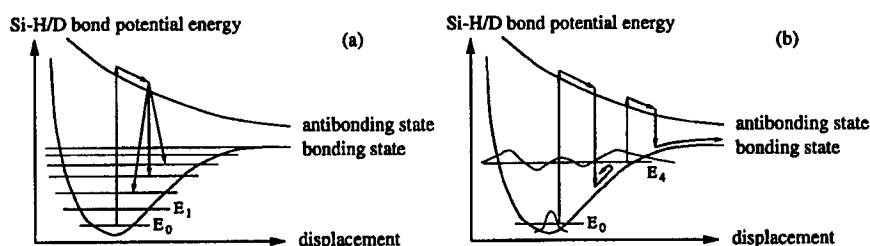


Figure 4. Possible hybrids of the coupling mechanisms of Fig. 1. (a) Excitations of the bond electron that fail to produce desorption directly, contribute to bond heating. (b) Both the energy required for excitation of the bond electron to the antibonding state and the required accumulation of kinetic energy during the lifetime of the antibonding state, are reduced for the higher-energy vibrational states occupied as the bond heats.

higher-bias operation within the same lifetime constraints. Indeed, D-passivated MOSFETs would be essentially immune to desorption via bond-heating provided that the large isotope effect identified for this process in the STM system remains in the MOS system.

It is also possible in the MOS environment that desorption could occur via a hybrid of these two mechanisms, possibly as depicted in Fig. 4. In Fig. 4(a), excitations of the bond electron that fail to produce desorption directly contribute to bond heating, possibly contributing multiple phonons of energy per scattering event, in contrast to direct excitation of the vibrational mode by channel hot carriers. In Fig. 4(b) excitation of the vibrational modes makes both excitation of the bond electron by hot carriers and subsequent desorption more probable, by reducing the energy requirement for excitation to the antibonding state and by requiring the accumulation of less kinetic energy during the lifetime of the antibonding state, respectively. However the coupling models of Fig. 4 must be considered only speculative. STM measurements have not probed the conditions of high electron current and high electron energy simultaneously to provide an experimental reference, and not only are the details of each desorption mechanism not yet fully understood theoretically, but at this time it is not even clear that desorption occurs by the same path for the two desorption mechanisms. To address such issues a first-principles analysis of Si-H/D bonds in the MOS environment is required.

4. First-principles analysis of desorption at the Si/SiO₂ interface

As discussed in the previous sections, hot carriers provide a source of energy to allow desorption of H/D from Si-H/D bonds at the Si/SiO₂ interface. A desorbed H/D atom presumably leaves behind a silicon dangling bond (Si_{db}) and thereby

creates an electronic defect. The dissociation of Si-H bonds can be described by the equation,



where H^* , the final configuration for the hydrogen atom, depends on the local chemistry and is largely unknown. If the final configuration is close to H in free space then the dissociation barrier will be at least the bonding energy. However, the dissociation barriers can be much lower for Si-H bonds at the Si/SiO₂ interface.

To quantify H/D desorption and transistor lifetime predictions necessitates an understanding of the atomistic mechanisms and associated energies for processes described by Eq. (6) for the Si/SiO₂ interface. Several relevant configurations have been investigated with first-principles total energy calculations.^{12,13} The present calculations are based on density functional theory within the local density approximation. We have investigated the energetics of Si-H bond dissociation within two frameworks. First, a periodic supercell model of an isolated Si-H bond in bulk crystalline silicon was examined. These results were verified by examining a large cluster model of an Si-H bond at the Si(111)/SiO₂ interface. A more detailed discussion of these calculations will be forthcoming.¹²

Two adiabatic dissociation paths for H and the energetics for some of the relevant configurations are presented in Fig. 5. Note that H and D are electronically identical, and so, therefore, are their bonding energies; the isotope effects result from differing dynamic properties of the bonds due to the differing masses of the isotopes. The zero of energy is set at the energy of H passivating an

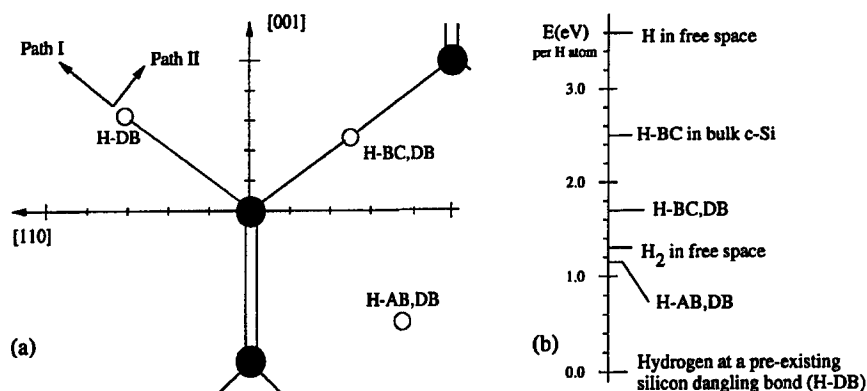


Figure 5. Competing dissociation paths and energetics for relevant configurations as determined from first-principles total energy calculations. (a) Local configuration for an isolated Si-H bond in silicon along with two competing dissociation paths, where the larger filled circles represent silicon atoms and the smaller open circles represent hydrogen atoms. (b) Total energies per H atom along the dissociation paths.

isolated dangling bond in bulk crystalline silicon. The energy of this Si-H bond should be similar to such bonds at the Si/SiO₂ interface and in bulk amorphous SiO₂.

Given a sufficiently large void above an Si-H bond, the dissociation along Path I of Fig. 5(a) will end in the formation of a silicon dangling bond and a neutral H in free space with the corresponding desorption energy of $E_d \sim 3.6$ eV as indicated in Fig. 5(b). Of course, at the Si-SiO₂ interface, the local chemistry above the Si-H bond may alter the desorption barrier. For our model of an Si-H bond at the Si(111)/SiO₂ interface, the dissociated H atom does not interact strongly with the oxide and no significant reduction of the dissociation barrier occurs. Upon examining the electronic structure of H along Path I, after a displacement of 1.0 Å at a relative total energy of 1.6 eV, a localized electronic state was found to form near the band edges. Thus, before full dissociation occurs in the neutral charge state, the H complex can accept free electrons, significantly reducing further barriers to dissociation.

Dissociation along Path II can lead the hydrogen to the anti-bonding (AB) site behind the dangling bond site. The hydrogen initially moves perpendicular to the bond direction, as indicated in Fig. 5(a), then it curves toward the neighboring bond-center site which we label (BC, DB) to distinguish it from the bond-center site far from any defects. The (BC, DB) site constitutes a local maxima. Finally, the hydrogen will rest at the (AB, DB) site. The maximum at the (BC, DB) site is rather flat. For instance, sites within 0.5 Å of the (BC, DB) site have roughly the same energy which is ~ 1.8 eV higher than H-DB. Also, the (H-BC, DB) configuration has an associated localized electronic state near both band edges. The defect wavefunctions are localized on the H atom and the silicon dangling bond. While near the BC site, the hydrogen complex can also accept free carriers. The hydrogen at the (AB, DB) site is lower in energy than at the (BC, DB) site by 0.5 eV. While in the stable (H-AB, DB) configuration, the H is electrically active and free carriers can be accepted which would greatly reduce further barriers to dissociation.

The final configuration for H desorption along Path II will involve H at a bulk interstitial site. As indicated in Fig. 5(b), the energy of the neutral H-BC interstitial is only 2.5 eV higher in energy than the H-DB. Once in the BC site, the H atom is mobile in MOS transistors operating at room temperature. Interstitial hydrogen will subsequently lower its energy, e.g., by binding to other defects or by forming an H₂ molecule. At surfaces or interfaces with open materials such as SiO₂, H₂ can easily escape from the material.

An important conclusion from this investigation is that the Si-H dissociation barrier can be significantly less than the Si-H binding energy, particularly if free carriers are present. An examination of Si-H dissociation in models of the Si(111)/SiO₂ interface confirm the above conclusions. If the primary hydrogen passivated defects can be characterized as dangling bonds, then these results should apply to Si-H bonds at the technologically more relevant Si(100)/SiO₂ interface.

5. Conclusion

Both the need for and challenges to development of a first-principles understanding of hot-carrier degradation in MOS devices have been addressed. A close correspondence between STM hydrogen and deuterium desorption measurements on silicon surfaces and the hot-carrier degradation and aging of MOSFETs has been demonstrated, providing insights that could lead the way to a first-principles understanding of at least one important degradation mechanism in MOS devices. The discussion in Sections 3 and 4 illustrates the challenges to and rudimentary steps toward extending the basic understanding of H/D desorption provided by the STM experiments to a first-principles-based modeling capability for simulation of hot-carrier degradation in MOS devices. In particular, the wide variations of desorption energies depending on the desorption pathway, the complications due to the interface with silicon dioxide, the details of hot-carrier interactions with the Si-H bond, and the possibility of mixing of the two basic desorption mechanism in the MOS environment, indicate that there is much work yet to be performed before it will be possible to quantitatively understand the isotope effect in transistors and to predict transistor lifetimes from first principles. Nevertheless, because the lifetimes of devices under normal operating conditions can be directly verified only years and perhaps several product generations after devices go to market, the incentive for development of first-principles-based degradation models to augment existing empirical models remains clear.

6. Acknowledgments

K. H. and L. F. R. were supported by ONR and the ARO. B. T. acknowledges support from the National Center for Supercomputing Applications, the DOE (DEFG 02-96-ER45439) and Stanford University (DARPA contract DABT63-94-C-0055). J. L. was supported by ONR.

References

1. R. W. Keys, "The battle with hot electrons," *Physics World* 9, 26 (1996).
2. K. R. Mistry and B. Doyle, "AC versus dc hot-carrier degradation in *n*-channel MOSFETs," *IEEE Trans. Electron Dev.* 40, 96 (1993).
3. C. H. Hu, S. C. Tam, F.-C. Hsu, *et al.*, "Hot-electron-induced MOSFET degradation — model, monitor, and improvement," *IEEE Trans. Electron Dev.* 32, 375 (1985) and references therein.
4. E. Takeda, C. Y. Yang, and A. Miura-Hamada, *Hot Carrier Effects in MOS Devices*, New York: Academic Press, 1995, pp. 88-90.
5. T.-C. Shen, C. Wang, G. C. Abeln, *et al.*, "Atomic-scale desorption through electronic and vibrational excitation mechanisms," *Science* 268, 1590 (1995).

6. E. Foley, A. F. Kam, J. W. Lyding, and Ph. Avouris, "Cryogenic UHV-STM study of hydrogen and deuterium desorption from Si(100)," *Phys. Rev. Lett.* **80**, 1336 (1998).
7. D. Menzel and R. Gomer, "Desorption from surfaces by slow-electron impact," *J. Chem. Phys.* **41**, 331 (1964).
8. C. G. Van de Walle and W. B. Jackson, "Comment on 'Reduction of hot-electron degradation in metal oxide semiconductor transistors by deuterium processing'," *Appl. Phys. Lett.* **69**, 2441 (1996).
9. J. W. Lyding, K. Hess, and I. C. Kizilyalli, "Reduction of hot-electron degradation in metal oxide semiconductor transistors by deuterium processing," *Appl. Phys. Lett.* **68** 2526 (1996);
I. C. Kizilyalli, J. W. Lyding, and K. Hess, "Deuterium post-metal annealing of MOSFETs for improved hot carrier reliability," *IEEE Electron Dev. Lett.* **18**, 81 (1997)
10. K. Hess, L. F. Register, B. Tuttle, J. Lyding, and I. C. Kizilyalli, "Impact of nanostructure research on conventional solid state electronics: the giant isotope effect in hydrogen desorption and CMOS lifetime," in: *Proc. 10th Intern. Winter School: New Frontiers in Low-Dimensional Physics*, Mauterndorf, Austria, 1998, to be published in *Physica E*.
11. A. Abramo, L. Baudry, R. Brunetti, *et al.*, "A comparison of numerical solutions of the Boltzmann transport equation for high-energy electron transport in silicon," *IEEE Trans. Electron Dev.* **41**, 1646 (1994).
12. B. Tuttle and C. G. Van de Walle, "First-principles study of Si-H dissociation in silicon," submitted to *Phys. Rev. B* (1998).
13. C. G. Van de Walle, "Energies of various configurations of hydrogen in silicon," *Phys. Rev. B* **49**, 4579 (1994).

2 Beyond CMOS: SOI, Heterostructures, Thin films

This chapter covers new technologies, primarily from the standpoint of a device physicist. The old dream of silicon-on-insulator (SOI) has received strong recent boosts from material technologists and can now be considered for possible inclusion in the mainstream silicon technology. Silicon-germanium heterostructures have also received a strong boost from their successful utilization in heterojunction bipolar transistors. These technologies have matured to the point that they can be contemplated for inclusion in the VLSI repertoire. Is there a role for SiGe in submicron CMOS? Is there an opening for heterojunction bipolars, either in VLSI or communications? Can one make use of ballistic and even quantum effects to achieve faster circuits? And then we make a leap to ultrahigh frequency computing, involving superconducting circuits and novel "petaflop" designs based on the RSFQ logic.

The chapter opens with two papers discussing SOI technologies. In the past, the focus of many efforts in the SOI area had been on the isolation of individual devices in integrated circuits and for niche applications, such as radiation-hard and, perhaps, ultra-fast circuits. Today, the much wider pull of low-power low-voltage electronics is driving these efforts. Should everything become SOI? If so, this switch should also fundamentally change the prospects for incorporation of other non-conventional elements into silicon VLSI, such as heterostructure devices (not necessarily silicon-based) and quantum devices (not necessarily semiconductor!).

The other important pull for non-conventional electronic elements comes from the renewed demand for high-performance computing. As a recent report of the U.S. President's Information Technology Advisory Committee notes: "powerful computers have laid the foundation for societal advances such as designing cancer-fighting drugs and understanding the causes of pollution. As such, the continued advancement of high-end computing will be needed in areas it has not touched in the past, such as supporting large World Wide Web sites and simulating natural crises". Significant new investments are expected in this area with "the goal of attaining sustained petaops/petaflops on real applications by 2010".

SOI Technology: Renaissance or Science Fiction?

S. Cristoloveanu

Laboratoire de Physique des Composants à Semiconducteurs (UMR 5531)

ENSERG, B. P. 257, 38016 Grenoble Cedex 1, France

1. Introduction

The status of SOI technologies is briefly reviewed in terms of material synthesis, device architecture, performance, and trends. The merits and weaknesses of fully- and partially-depleted SOI MOSFETs are critically addressed. It is concluded that SOI will play a significant role in the microelectronics future by extending the frontiers of bulk silicon, if persisting problems can be rapidly solved.

Three distinct periods can be identified in silicon-on-insulator (SOI) technologies. During the "Stone Age" in the 1970's, expensive sapphire substrates were used for epitaxial growth of thin silicon films on which radiation-hard SOS devices were processed — see Fig. 1(a). Since then, new SOI structures have been conceived with the aim of dielectrically separating, using a buried oxide, the active device volume from the detrimental influence of the silicon substrate — see Fig. 1(b).

It was repeatedly claimed (and often admitted) that SOI devices could successfully compete with their bulk silicon counterparts envisioned for after-the-next generation ($n + 2$) of integrated circuits. But, for about two decades, many ($n + 2$) generations were born ... and lived happily, without SOI entering the commercial arena.

More recently (1995-98), impressive SOI circuits have been demonstrated, with direct impact on mainstream microelectronics: 0.5-V—200-MHz micro-processor,¹ 4-Mbit SRAM,² 16-Mbit—1-Gbit DRAM,³ and others.^{4,5} Alternatively, devices for specific SOI niches (high temperature, high power, radiation, sensors, etc.) have also been fabricated. They make use of special SOI features, including the possibility to (i) combine bulk Si and SOI on a single chip

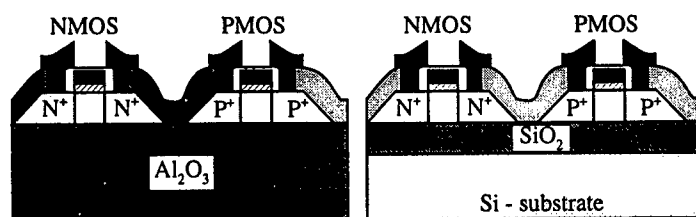


Figure 1. Configuration of MOS transistors on silicon-on-sapphire (SOS) (left) and silicon-on-insulator (SOI) (right) substrates.

(Fig. 2(a)), (ii) implement additional gates in the buried oxide (Fig. 2(b)), (iii) achieve perfectly controlled, thin membranes by using the buried oxide as an etch-stop layer (Fig. 2(c)), or (iv) adjust the thickness of the Si overlay and buried oxide.⁴ However, the real chance for SOI to become competitive is in the field of low-voltage/low-power (LV/LP) integrated circuits. This opportunity is why the recent period can be referred to as the "Renaissance".

The future trends of SOI-based microelectronics will depend heavily on the penetration rate of LV/LP SOI circuits into the marketplace. For the next millennium, SOI also offers the opportunity to integrate highly innovative devices that may extend the present frontiers of the CMOS downscaling. This period, referred to as "Science Fiction", includes double-gate or surrounding gate transistors (Fig. 2(d)), quantum devices, and three-dimensional (3D) stacked circuits.^{4,5} Science fiction is here used in a positive generic sense, which implies that such devices have already been demonstrated in terms of technology and functionality ... but most people still do not believe that they can actually operate.

The aim of this paper is to give a flavor of the state of the art of SOI technology and raise the most critical and provocative questions related to its advent.

2. Key advantages of SOI

SOI circuits consist of single-device islands dielectrically isolated from each other and from the underlying substrate. The lateral isolation offers more compact design and technology than in bulk silicon: there is no need of wells or interdevice trenches. On the other hand, the vertical isolation allows erasing the word *latch-up* from the SOI dictionary.

Since the source and drain regions extend to the buried oxide, the junction area is minimized, implying reduced junction capacitances and leakage currents. These reductions further translate into improved speed, lower power dissipation, and extended operation temperature ($\geq 300^\circ\text{C}$).

SOI devices are also less affected by short-channel effects because of the limited extension of the drain and source regions. Besides the outstanding tolerance of transient radiation effects, SOI MOSFETs experience a lower peak electric field than in bulk Si and are potentially more immune to hot carrier damage. In the domain of LV/LP transistors, where only a small gate voltage interval separates the off and on states, SOI brings the possibility of achieving a quasi-ideal subthreshold slope (60 mV/decade at room temperature), hence a threshold voltage below 0.3 V. This ideal situation occurs in fully-depleted SOI MOSFETs, where the depletion region covers the whole transistor body. Since the depletion charge remains constant, a better coupling develops between the gate bias and the inversion charge, leading to enhanced drain current.

The overall advantage of SOI is unequivocally summarized by the results of SOI against bulk Si contests. While operation at similar *voltage* consistently shows about a 30% increase in performance, operation at similar *low-power* dissipation yields as much as 300% performance gain in SOI. An SOI legend

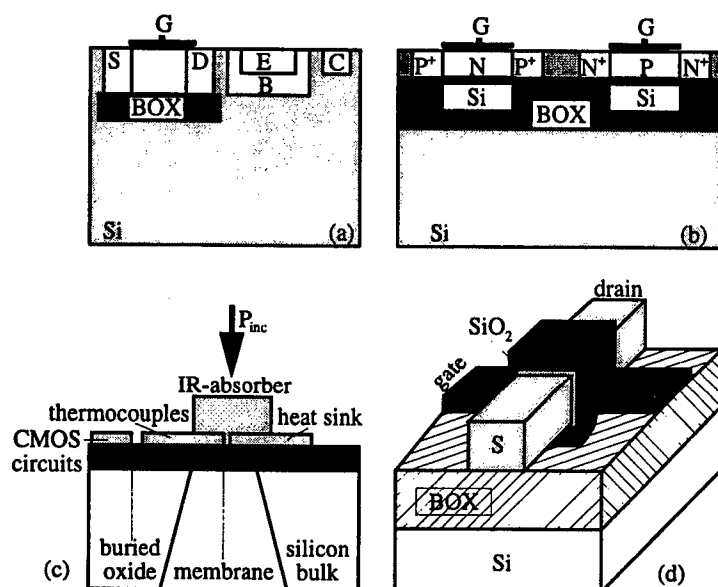


Figure 2. Examples of innovative SOI devices: (a) combined bipolar (or high power) bulk Si transistor with a SOI low-voltage control circuit, (b) dual-gate transistors, (c) infrared sensor, and (d) gate-all-around (GAA) MOSFET.

states that SOI circuits of generation n and bulk Si circuits from the *next* generation ($n + 1$) perform comparably.

Another view of SOI merits can be taken merely from the bulk Si technology, which desperately tries to mimic a number of features that are natural in SOI. For example, the double-gate configuration is reproduced by processing surround gate vertical MOSFETs on bulk Si, whereas full depletion is approached by tailoring a low-high step doping.

Of course, the foregoing enthusiastic list of SOI advantages originates from SOI advocates. This enthusiasm has not perturbed the fantastic progress of bulk Si technology so far and will not prevent two natural questions from being asked:

- Is this advantage enough for SOI to succeed?
- Is SOI more than an educational hobby for device physicists who are terribly bored with bulk Si and single-gate MOSFETs?

3. Availability of SOI wafers

A variety of techniques, more or less effective, are available for the synthesis of SOI wafers.⁴ Silicon-on-sapphire (SOS) has recently undergone some degree of

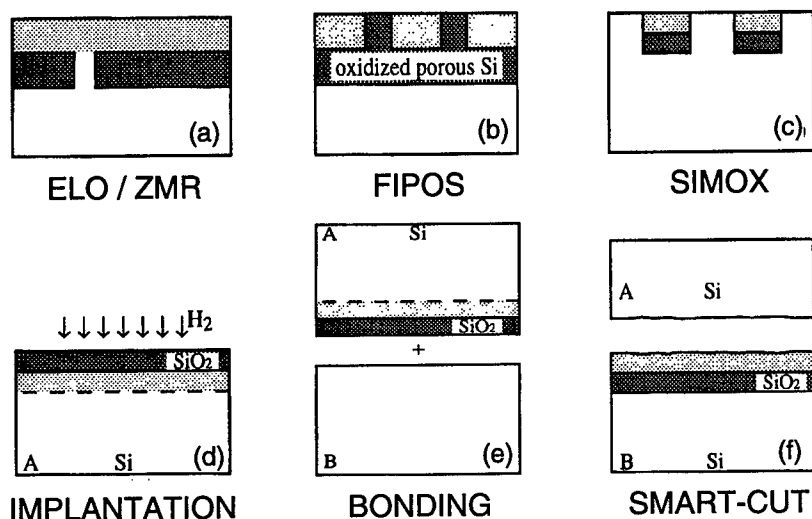


Figure 3. Family picture of SOI materials.

rejuvenation (larger wafers and thinner films with improved crystal quality). Thanks to the "infinite" thickness of the insulator, SOS looks promising for the integration of RF circuits.

The *epitaxial lateral overgrowth* (ELO) method consists of growing a single-crystal Si film on a seeded (and, sometimes, patterned) oxide (Fig. 3(a)). Since the epitaxial growth proceeds in both lateral and vertical directions, the ELO process requires a post-epitaxy thinning step. Alternatively, polysilicon can be deposited directly on SiO₂. Subsequently, *zone melting recrystallization* (ZMR) is achieved by scanning a high-energy source (lamps, lasers, strip heaters) across the wafer. The ZMR process, which can be seeded or unseeded, is basically limited by the lateral extension of single-crystal regions that are free from sub-grain boundaries and associated defects. ELO and ZMR are competing techniques for the integration of 3-D stacked circuits.

The FIPOS method (*full isolation by porous oxidized silicon*) makes use of the very large surface-to-volume ratio ($10^3 \text{ cm}^2 \text{ per cm}^3$) of porous silicon, which is therefore subject to selective oxidation (Fig. 3(b)). The critical step is the conversion of predefined regions of the Si wafer into porous silicon, via appropriate doping and anodic reaction. FIPOS may bring some light into Si technology because there are prospects, at least in theory, for combining electroluminescent porous Si devices with fast SOI-CMOS circuits.

Wafer bonding (WB) and etchback stands as a more mature SOI technology. An oxidized wafer is mated to another SOI wafer (Fig. 3(e)). The challenge is to dramatically thin down one side of the bonded structure in order to achieve the targeted thickness of the silicon film. Etch stop layers can be prepared by differential doping or porous silicon. A revolutionary thinning process

(UNIBOND) uses the deep implantation of hydrogen (Fig. 3(d)).⁶ After bonding and subsequent annealing, the hydrogen-induced microcavities coalesce and the two wafers separate (by the so-called *smart-cut* mechanism, Fig. 3(f)) leaving an SOI structure. The process is completed by touch-polishing to erase the surface roughness.

The extraordinary potential of the *smart-cut* approach comes from several distinct advantages: (i) the etchback step is avoided, (ii) since the second wafer (Fig. 3(f)) can be recycled, UNIBOND is a single-wafer process, (iii) only conventional equipment is needed for mass production, (iv) inexpensive 12" wafers are manufacturable, and (v) the thickness of the silicon film and/or buried oxide can be adjusted to match most device configurations (ultra-thin CMOS or thick-film power transistors and sensors). In addition, the process is adaptable to a variety of materials: SiC or III-V compounds on insulator, silicon on diamond, etc. *Smart-cut* can be used to transfer already fabricated bulk Si CMOS circuits on glass or other substrates.

Prior to the advent of UNIBOND, the dominant SOI technology was SIMOX (separation by implantation of oxygen). The deep implantation of high doses of oxygen into silicon naturally results in the synthesis of a buried oxide (BOX). The family of SIMOX structures includes: thick Si films (standard dose, $< 0.4 \mu\text{m}$ thick BOX), thin Si films (low dose, $< 0.1 \mu\text{m}$ thick BOX), interrupted oxides (Fig. 3(c)), laterally isolated islands, and double SIMOX, where the Si layer sandwiched between two oxides can serve for interconnects, wave guiding, additional gates, or electric shielding.

4. Fully or partially depleted SOI MOSFETs?

In SOI MOSFETs (Fig. 1(b)) inversion channels can be achieved either at the front Si—SiO₂ interface (via gate modulation V_{G1}) or at the back interface (via substrate, backgate bias V_{G2}). The typical thin-film SOI transistor is *fully depleted*, which

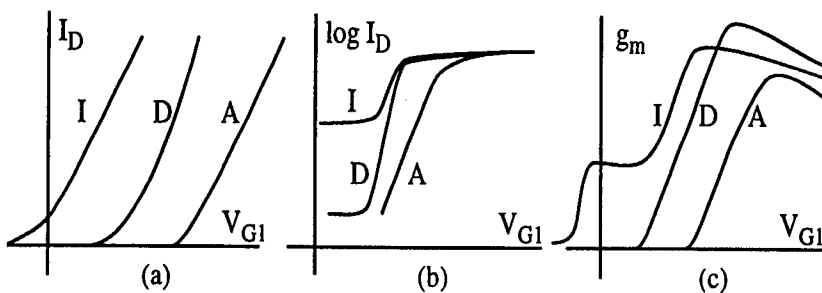


Figure 4. Generic front-channel characteristics of a fully-depleted SOI MOSFET for accumulation (A), depletion (D), and inversion (I) at the back interface: (a) strong inversion, (b) weak inversion, (c) transconductance.

means that the depletion region covers the whole film. In other words, the depletion region cannot extend according to the gate bias, hence the front and back surface potentials become interrelated. The coupling factor is roughly equal to the thickness ratio between gate oxide and buried oxide. The electrical characteristics of one channel vary remarkably with the bias applied to the opposite gate. Due to *interface coupling*, the front gate measurements are all reflective of the bias and quality of the buried oxide and interface. *Defect coupling* is observed as an apparent degradation of the front channel properties that is actually induced by the buried oxide damage (for example, after hot-carrier injection into the BOX).

Totally new $I_D(V_G)$ relations apply to fully-depleted SOI MOSFETs, whose complex behavior is controlled by both gate biases.⁷ The main characteristics of the front-channel transistor are schematically illustrated in Fig. 4, for three typical bias conditions of the back interface: accumulation, depletion and inversion. The lateral shift of the $I_D(V_G)$ curves (Fig. 4(a)) is due to the linear variation of the front-channel threshold voltage with back gate bias (i.e. potential coupling). The subthreshold slope (Fig. 4(b)) is very steep for depletion at the back interface, when the transconductance peak is also a maximum (Fig. 4(c)). The latter effect is explained by the reduction of the vertical field and series resistances. The distortion of the transconductance curve reflects the possible activation of the back channel before the inversion charge is built up at the front channel.

In thin and low-doped films, the simultaneous activation of front and back channels causes by continuity (i.e. charge coupling) the onset of *volume inversion*.⁸ Unknown in bulk Si, this effect enables the inversion charge to cover the whole film and results in increased current drive. Double-gate MOSFETs also benefit from reduced short-channel effects (punchthrough, drain-induced barrier lowering, etc), and are therefore very attractive for downscaling below 30-nm gate length.

In partially-depleted SOI MOSFETs, where the depletion charge controlled by one or both gates does not extend from one interface to the other, a neutral region exists. The interface coupling effects are therefore disabled. When the body is grounded (via independent body contacts or body—source ties), partially-depleted SOI MOSFETs behave very much like bulk Si transistors and most of the standard $I_D(V_G, V_D)$ equations and design concepts apply. Otherwise, detrimental so-called *floating-body* effects arise.

For example, majority carriers generated by impact ionization collect in the transistor body: The body potential is raised, which in turn lowers the threshold voltage. This feedback induces a kink in the $I_D(V_D)$ characteristics, shown in Fig. 5(a), which is annoying in analog circuits. In weak inversion and for high drain bias, a similar positive feedback (increased inversion charge \rightarrow more impact ionization \rightarrow body charging \rightarrow threshold voltage reduction) gives rise to negative resistance regions, hysteresis in $\log I_D(V_G)$ curves, and eventually latch (loss of gate control, Fig. 5(b)).

The floating body is also responsible for the occurrence of transient effects. A current *overshoot* is observed as the gate is turned on — see Fig. 5(c). Majority carriers are expelled from the depletion region and collect in the neutral body increasing the potential. Equilibrium is reached through carrier recombination,

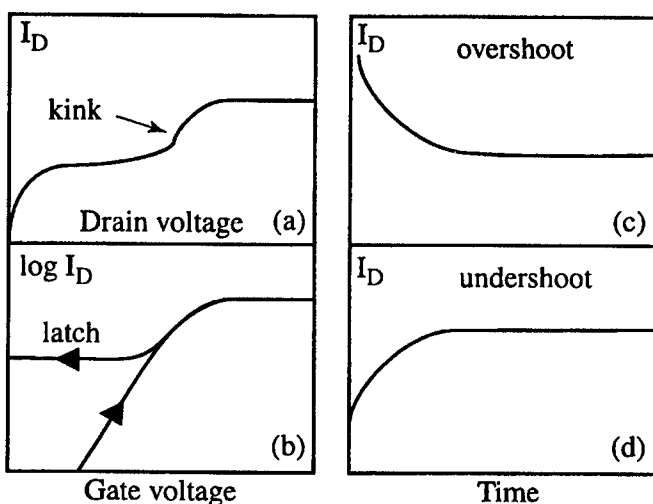


Figure 5. Typical characteristics of partially-depleted SOI MOS transistors: (a) kink in $I_D(V_D)$ curves, (b) latch in $I_D(V_G)$ curves, (c) current overshoot, (d) current undershoot.

which removes the excess majority carriers, making the drain current decrease gradually with time. A reciprocal *undershoot* occurs when the gate is switched from strong to weak inversion: the current increases with time (Fig. 5(d)) as the majority carrier generation allows the depletion depth to shrink.

An obvious solution to alleviate floating-body effects is to sacrifice some space for designing body contacts. However, in ultra-thin films with large sheet resistance, the body contacts are far from being ideal; their resistance does not allow the body to be perfectly grounded and may also generate additional noise. Thus, even in microelectronics, a floating body is still preferable to an incompetent body contact.

An exciting partially-depleted device, the dynamic-threshold DT-MOS transistor, is achieved by interconnecting the gate and the body. As the gate voltage increases in weak inversion, the concomitant rise in body potential makes the threshold voltage decrease. DT-MOSFETs exhibit very attractive features (perfect gate-charge coupling, maximum subthreshold slope, enhanced current) for low-voltage/low-power circuits.

In both fully and partially depleted MOSFETs with submicron length, the source-body junction can easily be turned on. This turn-on activates the inherent lateral bipolar transistor, the contribution of which is positive (extra current flow in the body) or negative (premature breakdown, Fig. 6(b)). Another problem is the *self-heating*, exacerbated by the poor thermal conductivity of the surrounding SiO_2 layers. Self-heating is responsible for mobility degradation, threshold voltage shift, and negative differential conductance (Fig. 6(a)).

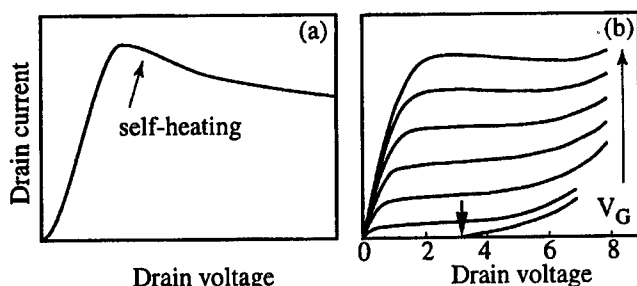


Figure 6. Self-heating effect (a) and early breakdown (b) in short-channel SOI MOSFETs.

5. SOI challenges

Although SOI is already pretty mature, there are still serious challenges in various domains: fundamental and device physics, technology, device modeling, and circuit design. The minimum dimensions so far achieved for SOI MOSFETs are: 70 nm length, 10 nm width (quantum wires), and 1-2 nm thickness. When these features are combined in a single transistor, the body volume ($\leq 10^{-18} \text{ cm}^3$!) will contain 10^4 – 10^5 silicon atoms and 0–1 defects. The body doping will be provided by a single impurity, whose location may become important. On the other hand, quantum transport phenomena are already being observed in ultra-thin SOI transistors. It is clear that new physical concepts, ideas, and modeling tools will be needed to account for minimum-size effects and use them.

On the technology side, a primary challenge is the mass production of SOI wafers with large diameter (≥ 12 "), low defect content, and reasonable cost. The thickness uniformity of the silicon layer is especially important for fully depleted MOSFETs. A number of SOI technologies will probably not survive, except in history books.

Appropriate characterization techniques, imported from other semiconductors or entirely conceived for SOI, are demanded. A technique unique to SOI is the pseudo-MOS transistor (Ψ -MOSFET).⁹ Ironically, it behaves very much like the MOS device that Shockley attempted to prove 50 years ago, but he didn't have the chance to know about SOI at that time. The Si substrate is biased as a gate and induces a conduction channel (inversion or accumulation) at the film-oxide interface. Source and drain probes are used to measure the $I_D(V_G)$ characteristics. The Ψ -MOSFET does not require any processing, hence valuable information is immediately available on the quality of the film, interface and oxide, as well as electron and hole mobilities and lifetimes.

Full CMOS processing must address typical SOI requirements such as the series resistance reduction in ultra-thin MOSFETs (local oxidation, elevated source and drain structures, etc), and the lowering of the source-body barrier by source engineering. The best of SOI is certainly not achievable simply by using a

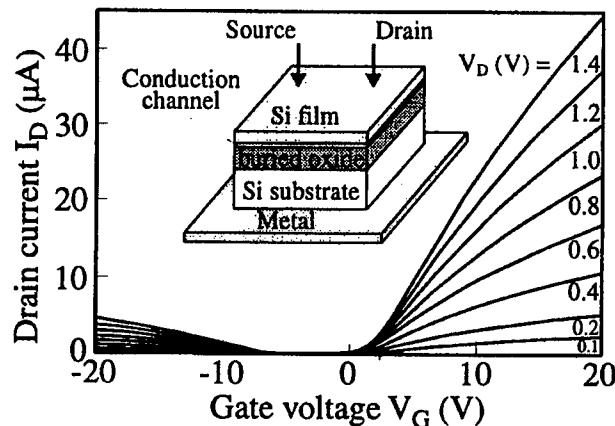


Figure 7. Pseudo-MOSFET transistor and $I_D(V_G)$ characteristics in SOI.

very good bulk Si technology. For example, double-gate SOI MOSFETs deserve special attention.

Partially depleted SOI MOSFETs are presently more user friendly to process engineers and designers, whereas fully-depleted transistors have more advanced capabilities. The latter should be domesticated to become more tolerant to short-channel effects, for example by incorporating a ground plane in the buried oxide.

Advanced modeling is demanded for correct transcription of the transistor behavior, including transient effects due to body charging and discharging, floating body mechanisms, bipolar transistor, dual gate operation, quantum effects, self-heating, and short-channel limitations. Based on such physical models, compact models should then be conceived for simulation and design purposes.

In particular, SOI does need SOI-dedicated CAD libraries, which stand as a huge task and which, in turn, will guarantee that the advantages and peculiar constraints of SOI devices are properly taken into account. The optimum configuration of memories, microprocessors, DSP, *etc.* will most likely be different in SOI compared to bulk Si. Indeed, SOI can afford to combine fully and partially depleted, low and high power, and DT-MOSFETs in a single chip.

The key challenge is mostly educational and strategic, oriented to overcoming the bulk Si monocultural barrier. Designers, process engineers, and managers keep very busy loading the bulk Si machine. Nevertheless, they should take a long moment for contemplating the assets of SOI technology. In doing so they will realize the immediate and long-term benefits offered in terms of performance and scaling extensions.

SOI is not really different, it is just another (good) taste of silicon ...

References

1. T. Fuse, Y. Oowaki, T. Yamada, *et al.*, "A 0.5 V 200 MHz 1 stage 32 bit ALU using a body bias controlled SOI pass-gate logic," *ISSCC Tech. Digest* **40**, 286 (1997).
2. D. J. Schepis, F. Assaderaghi, D. S. Yee, *et al.*, "A 0.25 μm CMOS SOI technology and its application to 4 Mb SRAM," *IEDM Tech. Digest* (1997), p. 587.
3. Y.-H. Koh, M.-R. Oh, J.-W. Lee, *et al.*, "1 Gigabit SOI DRAM with fully bulk compatible process and body-contacted SOI MOSFET structure," *IEDM Tech. Digest* (1997), p. 579.
4. S. Cristoloveanu and S. S. Li, *Electrical Characterization of SOI Materials and Devices*, Norwell: Kluwer, 1995.
5. J.-P. Colinge, *SOI Technology: Materials to VLSI*, 2nd ed., Boston: Kluwer, 1997.
6. M. Bruel, "Silicon on insulator material technology," *Electron. Lett.* **31**, 1201 (1995).
7. H.-K. Lim and J. G. Fossum, "Threshold voltage of thin-film silicon on insulator (SOI) MOSFETs," *IEEE Trans. Electron Dev.* **30**, 1244 (1983).
8. F. Balestra, S. Cristoloveanu, M. Bénachir, J. Brini, and T. Elewa, "Double-gate silicon on insulator transistor with volume inversion: a new device with greatly enhanced performance," *IEEE Electron Dev. Lett.* **8**, 410 (1987).
9. S. Cristoloveanu and S. Williams, "Point contact pseudo-MOSFET for in-situ characterization of as-grown silicon on insulator wafers," *IEEE Electron Dev. Lett.* **13**, 102 (1992).

Single- and Double-Gate SOI MOS Structures for Future ULSI: A Simulation Study

Claudio Fiegna

Dept. of Engineering, Univ. of Ferrara, Ferrara, I44100 Italy

Antonio Abramo and Enrico Sangiorgi

DIEGM, University of Udine, Udine, I33100 Italy

1. Introduction

Silicon-on-insulator MOS technology, originally proposed in order to circumvent the parasitic latch-up effect, has recently attracted great interest as a potentially viable solution for MOS scaling below $0.1\ \mu\text{m}$ due to superior control of short channel effects and lower leakage currents compared to bulk MOSFETs.¹ Fully depleted SOI devices featuring very thin silicon layers are usually indicated as the ideal choice in order to cope with short channel effects and to reduce parasitic capacitances, while keeping detrimental floating substrate effects (i.e. the kink effect) under control. Silicon thicknesses in the 10 nm range have been proposed and some experimental studies have been reported in the literature.²

Additional improvements can be obtained by adopting a symmetric double-gate SOI MOS structure that improves gate control over the channel potential and charge distribution, leading to further reduction of short channel effects.³ Furthermore, in the case of double-gate MOS (DGM), higher drain currents compared to a single gate SOI MOS (SGM) of same area were reported.^{4,5} The increase of the double-gate MOSFET's drain current is due, above all, to the formation of a double conducting channel close to the two Si/SiO₂ interfaces (see Fig. 1). This second conducting channel accounts for a factor of two difference with respect to a SGM biased at the same gate drive ($V_G - V_{TH}$). Additional gain in terms of transconductance and current drive (up to a factor of 2.5–3) were experimentally reported in Refs. 4, 5 and attributed to the inversion of the silicon region away from the two interfaces. In fact, a large amount of charge displaced from the interface would lead to higher carrier effective mobility due to reduced surface scattering. Furthermore, due to the interaction of the two symmetric gate electrodes, a larger control capacitance (i.e. a larger amount of inversion charge) may be expected in the DGM case compared to the SGM one when the two structures are biased at the same ($V_G - V_{TH}$). As a consequence, it is expected that $N_{S,DGM} > 2N_{S,SGM}$, where N_S is the inversion charge sheet concentration. Regarding the advantages of DGM over SGM, no agreement is found in the literature. In fact the results of simulations reported in Ref. 6 indicate that just a 10% gain in peak transconductance is to be expected in the DGM with respect to the SGM when compared at the same effective channel width (i.e. width of the SCM being twice

that of the DGM). As a consequence of this discrepancy, a debate has been started about these issues.^{7,8}

This article presents a study of the fundamental characteristics of SGM and DGM structures in order to clarify the following issues:

- the relevance of size-quantization in thin SOI structures and its effects on threshold voltage;
- the performance improvements in double gate structures, in particular i) the role played by volume inversion; and ii) how gate capacitance can be improved due to the interaction between the two gates;
- the dependence of low-field mobility on silicon thickness in SGM devices.

The paper is organized as follows: Section 2 presents the simulated structures, the physical models adopted, and the numerical techniques used in this study; Section 3 presents the simulation results; finally, Section 4 summarizes the conclusions of this work.

2. Simulated structures and methodology

In Fig. 1 a schematic picture of the simulated structures is sketched. In the DGM case, a thin silicon layer is included between two symmetric gate structures while in the SGM case, a much thicker back oxide is present. Since we are investigating a possible structure for MOS devices with gate length $L_G \leq 0.1 \mu\text{m}$, we assumed very thin gate oxide (t_{ox}) and silicon layer thicknesses (t_{Si}). In our simulations, we considered t_{Si} down to 5 nm, $t_{\text{ox}} = 3\text{--}7$ nm and the thickness of the back oxide of the SGM devices $t_{\text{box}} = 50$ nm. For the background silicon layers, p -doped silicon with $N_A = 10^{15}\text{--}10^{17} \text{ cm}^{-3}$ was assumed for both structures. The two gate electrodes of the DGM and the front gate of the SGM are assumed to be n -polysilicon, while in the SGM device, a grounded p -polysilicon gate mimics the effect of the silicon bulk.

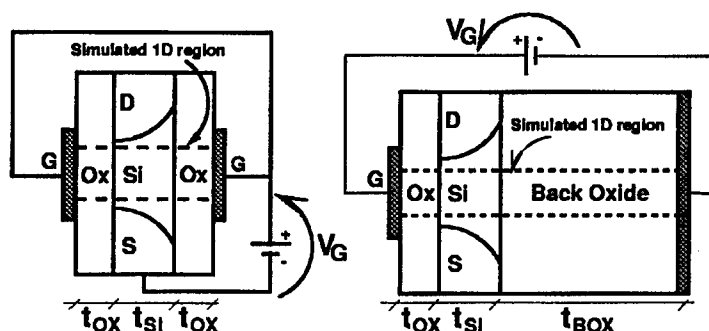


Figure 1. Schematic section of the simulated structures; left: double-gate MOS, right: single-gate MOS.

As for the silicon thickness, it is well known that its reduction diminishes short channel effects in both SGM⁹ and DGM.³ Furthermore, reducing t_{Si} enhances the effects of volume inversion in the DGM case.^{4,7} Previous experimental studies on single gate SOI reported a substantial mobility degradation when t_{Si} is reduced below 10 nm and attributed it to the mechanical stress of the silicon layer.² As we are particularly interested in volume inversion, we considered thinner t_{Si} in order to enhance volume inversion in the DGM structure and did not include the influence of the mechanical stress on mobility. Regarding the thickness of the SGM back oxide, the selected value is thinner than those obtainable by the conventional low-dose SIMOX SOI process (80100 nm),¹⁰ the main consequence being an increase of the transverse electric field and a decrease of short channel effects. The simulations were carried out by self-consistently solving the Poisson and Schrödinger equations.¹¹ The quasi-Fermi levels for electrons and holes are set within the whole simulation domain so as to reflect a bias condition with grounded source and drain. The envelope function equation (i.e. Schrödinger equation in the effective mass approximation) is solved to determine the eigenvalues and eigenvectors of the system. In this solution, the six-fold ellipsoidal symmetry is assumed for the silicon, with the usual values for longitudinal and transverse masses (0.19 and 0.915 in free electron mass units, respectively), together with a parabolic energy vs. k -vector dispersion relationship. A zero wavefunction boundary conditions is forced at the two oxide interfaces, neglecting penetration into the oxide.

Once the eigenstates are determined, the quantum electron density and the classical hole density are computed using Fermi-Dirac and Boltzmann statistics, respectively. Complete ionization of the dopants is also assumed. The fixed and mobile charge profiles are, then, provided to the non-linear Poisson solver that returns a consistent potential profile. Schrödinger and Poisson equations are then iteratively solved with a Gummel-like procedure.

3. Simulation Results

To check the procedure adopted for the simulation of floating-body devices, Figure 2 reports the potential energy, the quantized energy levels and electron concentration within the inversion layer calculated for a relatively thick DGM (t_{Si} large enough so that the silicon is never completely depleted) and a conventional bulk MOSFET with grounded body. The comparison is performed for the same amount of inversion charge in the bulk-MOSFET and in each of the two DGM's symmetric inversion layers. Potential energy profiles, energy levels and electron concentrations within the inversion layer perfectly overlap confirming that consistency is obtained with respect to the conventional bulk-MOSFET case.

- *Quantization effects in very thin SOI structures*

The relevance of quantization in ultra-thin SOI devices is demonstrated by Figs. 3 and 4. Figure 3 compares potential energy and quantized levels for a conventional bulk MOSFET and a SGM with 10 nm silicon thickness. Inversion

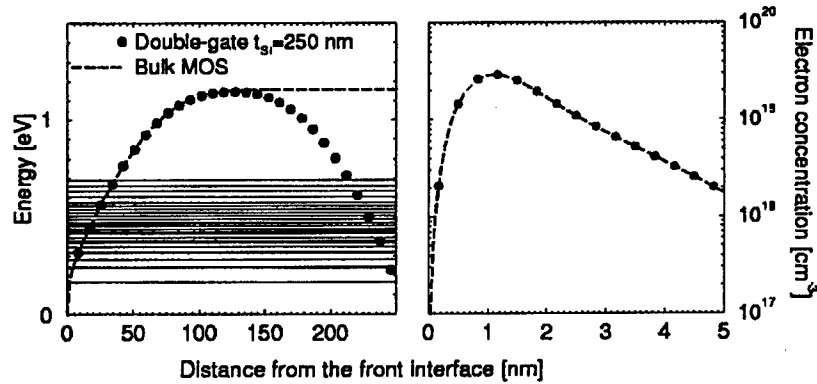


Figure 2. Potential energy, quantized energy levels and electron concentration within the inversion layer calculated for a non-fully depleted DGM and a bulk MOSFET with grounded body. Comparison is made for the same inversion charge in bulk MOS and within each of the DGM inversion layers ($6 \times 10^{12} \text{ cm}^{-3}$).

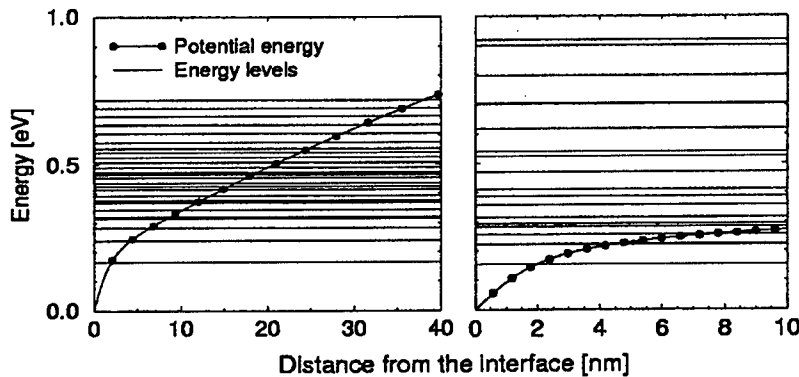


Figure 3. Potential energy, quantized energy levels within the inversion layer calculated for a bulk MOSFET with grounded body (right) and a fully depleted SGM with $t_{si} = 10$ nm (left). Comparison is made for the same inversion charge in both bulk MOS and DGM ($6 \times 10^{12} \text{ cm}^{-3}$; note that different x-axis scales are adopted in the two figures).

layer sheet density is the same in the two cases, but only the first few energy levels are almost coincident in the two structures. In the SGM case a much larger quantization occurs due to the presence of the two Si-SiO₂ interfaces. In the case of bulk MOSFETs and thicker SOI devices, the quantum-mechanical approach predicts an electron concentration displaced towards the depth of the device, while the potential profile is only marginally changed compared to the classical solution.

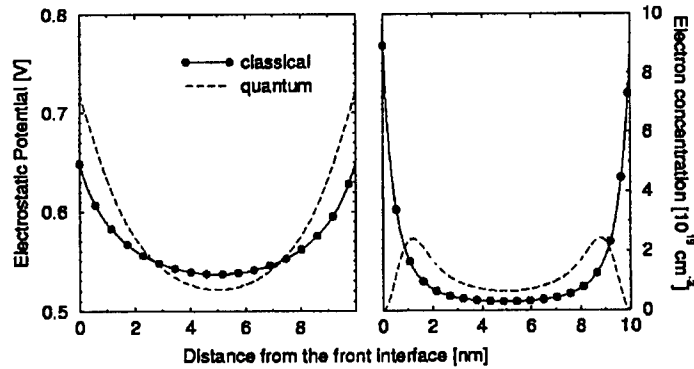


Figure 4. Potential and electron density along a one-dimensional DGM structure as calculated by classical calculations and by self-consistent solution of Schrödinger and Poisson equations. $N_A = 10^{17} \text{ cm}^{-3}$, $t_{ox} = 3 \text{ nm}$, $t_{Si} = 10 \text{ nm}$.

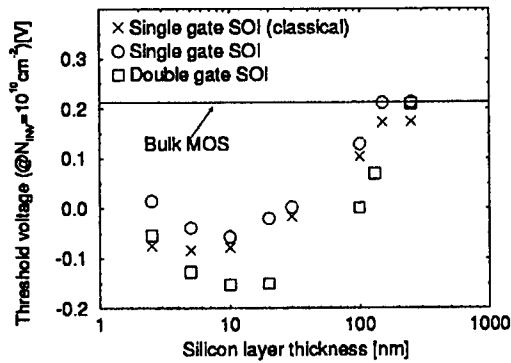


Figure 5. Threshold voltage vs. silicon layer thickness for the SGM (circles) and the DGM (squares) structures; solid lines refers to a bulk MOSFET with the same uniform doping concentration as the SOI ones. Crosses refer to classical calculations for the DGM. $N_A = 10^{17} \text{ cm}^{-3}$, $t_{ox} = 7 \text{ nm}$.

On the other hand, Figure 4 shows that the potential and electron distribution calculated for a DGM with $t_{Si} = 10 \text{ nm}$ by classical and quantum-mechanical models are quite different from each other within the whole structure due to substantial size-quantization. Therefore, we must conclude that the quantum-mechanical model is mandatory for studying such ultra-thin structures.

- *Threshold voltage dependence on silicon thickness and doping*

In this section threshold voltage dependencies of ultra-thin SGM and DGM are reported. Figure 5 reports the threshold voltage (calculated as the linear extrapolation of charge vs. gate voltage characteristic) as a function of silicon

thickness for both SGM and DGM with $N_A = 10^{17} \text{ cm}^{-3}$. Open circles refer to the SGM while squares refer to the DGM. As the silicon thickness is reduced below a critical value dependent on the doping concentration ($\sim 100 \text{ nm}$ for SCM and $\sim 200 \text{ nm}$ for the DGM when $N_A = 10^{17} \text{ cm}^{-3}$) the threshold voltage is reduced below the value corresponding to the conventional bulk MOSFET as a consequence of the reduction of depletion charge in strong inversion.

As the silicon thickness becomes small enough to produce substantial size quantization ($\sim 10 \text{ nm}$ for SGM) the threshold voltage starts increasing due to the reduction of the available density of states as a consequence of strong separation between the energy levels (see also Figure 3). The neglect of quantum-mechanical effects (crosses) leads to a negative threshold shift in the case of non-depleted silicon (consistent with well known results reported for bulk MOSFETs¹²) while in the limit of ultra thin silicon layer the threshold voltage is almost independent of t_{Si} , within the precision of threshold voltage extrapolation procedure.

- *Volume inversion in the DGM*

To enhance the effect of volume inversion, we considered extremely thin silicon layers (down to 5 nm). Figure 6 reports the electron concentration in a DGM structure biased at two different gate voltages above threshold. A maximum located at the center of the silicon film is obtained for bias points close to the threshold, while at higher V_G two separate inversion charge maxima are formed with a non-negligible concentration in the silicon volume. The effect of volume inversion vanishes rapidly as t_{Si} is increased towards more realistic values, leading to a reduction of the electron concentration in the middle of the silicon film. This effect is reported in Figure 6, which shows the electron concentration at the midpoint of the silicon layer normalized to the peak concentration close to the interfaces, as a function of t_{Si} and for two given gate drive voltages above threshold.

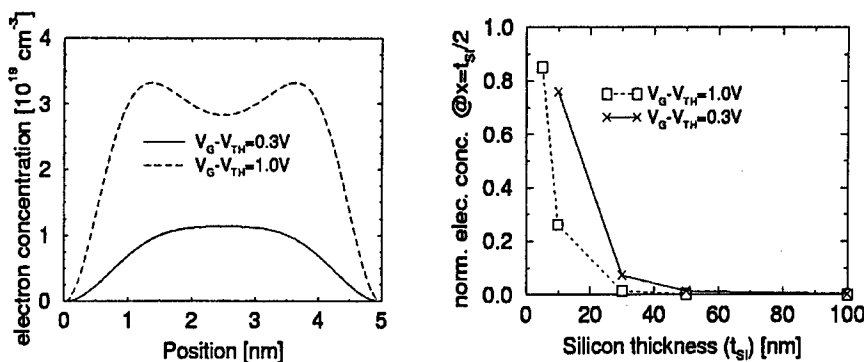


Figure 6. Left: electron concentration within the DGM structure for two different bias conditions above threshold; $N_A = 10^{17} \text{ cm}^{-3}$, $t_{\text{ox}} = 3 \text{ nm}$, $t_{\text{Si}} = 5 \text{ nm}$. Right: electron concentration evaluated at the middle of the silicon layer normalized by the maximum concentration as a function of silicon thickness; $N_A = 10^{17} \text{ cm}^{-3}$, $t_{\text{ox}} = 3 \text{ nm}$.

- *Comparison between DGM and SGM*

We have evaluated by simulation the main possible reasons for improved static characteristics of DGM devices.

For very thin silicon layers and low gate voltages (Fig. 6), the presence of a maximum of the electron concentration at the middle point, which is due to the interaction of the two gate fields, may be the sign of a larger inversion sheet density for the DGM compared to the SGM biased at the same gate drive. In particular, in the presence of substantial volume inversion, a larger inversion charge and a larger derivative with respect to the gate voltage (dQ_s/dV_G , $Q_s = eN_s$ being the inversion charge sheet density) could be expected in the DGM case compared to the SGM one. To further investigate this issue, we have performed a comparison between DGM and SGM capacitors with the same silicon thicknesses and biased above threshold. The results of simulations show that even when t_{Si} is comparable to the displacement of the charge peak from the interface, the increase of DGM inversion charge and capacitance occurs only for bias points close to the threshold voltage and it is almost negligible, while above threshold the two structures show coincident behaviors. This agreement is shown in Fig. 7, reporting the comparison between the dQ_s/dV_G of a DGM and a SGM structures with $t_{Si} = 5$ nm and the same effective width (i.e. the charge per unit width of the SGM is multiplied by a factor of two). From this result we may not expect a large improvement in the current and transconductance as a direct consequence of the increase of inversion charge due to the interaction between the two gates of the DGM structure. Another possible explanation for the larger current and transconductance in the DGM case compared to the SGM one relates to the different distribution of inversion charge in the two structures, when compared for the same inversion charge sheet density. In particular, in the DGM the presence of inversion charge far away from the interface may lead to an increased effective low-field mobility due to the reduced surface roughness scattering. In Fig. 8, the charge density profile of a DGM structure with $t_{Si} = 10$ nm and $t_{ox} = 3$ nm is

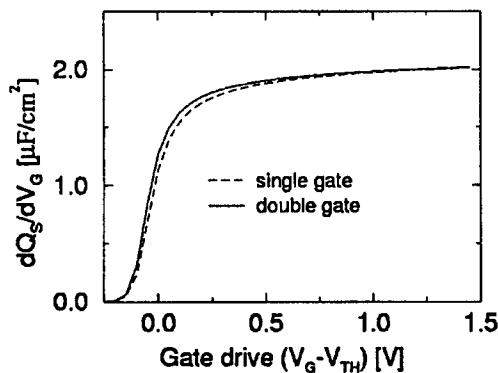


Figure 7. Gate capacitance vs. gate drive for ultrathin DGM and SGM structures. $N_A = 10^{17} \text{ cm}^{-3}$, $t_{ox} = 3 \text{ nm}$, $t_{Si} = 5 \text{ nm}$.

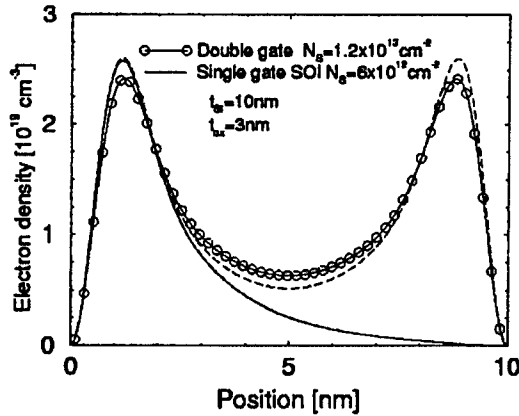


Figure 8. Electron concentration within the silicon layer of a SGM (solid line) and a DGM (circles and solid lines). The dashed line is the sum of two individual SGM charge profiles. $N_A = 10^{17} \text{ cm}^{-3}$, $t_{ox} = 3 \text{ nm}$, $t_{Si} = 10 \text{ nm}$.

reported and compared to that of a SGM with same silicon thickness. The two structures are biased at the same ($V_G - V_{TH}$) and the charge sheet density of the DGM case is twice that of the SGM. The DGM charge density profile is more displaced from the interface than the SGM one, and its value at the middle of the silicon layer is larger than what is obtained by summing the charge profiles of two individual SGMs (dashed line). However, such differences appear to be only marginal. Additional work is needed in order to check if significant improvements in the DGM effective mobility may be expected due to reduced surface scattering. In particular, a surface scattering model accounting for the simultaneous interaction of wavefunctions with both the front and the back interface is necessary and is still unavailable.

- *Low field mobility in ultra-thin SGM structures*

In this section we report the results of low-field mobility calculations performed by post-processing the results of Schrödinger-Poisson calculations in the framework of the relaxation time approximation. Only acoustic and optical phonons are accounted for in the calculations, while the effects of surface roughness are not included due to the lack of a model describing the interaction with both interfaces. Acoustic and optical phonon scattering are modeled using the model proposed in Ref. 13 for bulk silicon with the exception of the value for the acoustic deformation potential, here is assumed to be 11.5 eV consistent with previous studies of mobility in Si inversion layers.¹⁴

Figure 9 reports the calculated low-field mobility as a function of the effective field (average field within the silicon layer weighted by carrier concentration) for SGM structures with different thicknesses and $N_A = 10^{15} \text{ cm}^{-3}$. It is interesting to notice a nontrivial dependence of mobility on silicon thickness. For relatively

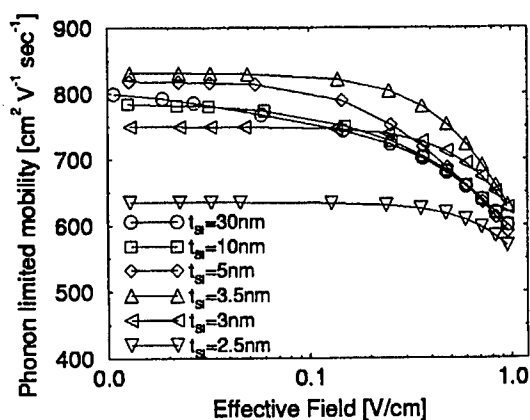


Figure 9. Low-field mobility vs. effective field for SGM structures with different Si layer thicknesses. $N_A = 10^{15} \text{ cm}^{-3}$, $t_{ox} = 3 \text{ nm}$.

thick silicon layers, low field mobility is comparable to that of bulk MOSFETs. However, as t_{Si} is scaled below $\sim 10 \text{ nm}$, the separation between energy levels increases due to size-quantization (Fig. 3) and, therefore, the available density of states for scattering is reduced, increasing the low-field mobility. A maximum is reached for $t_{Si} \approx 3.5 \text{ nm}$, while further reducing t_{Si} leads to rapidly decreasing mobility due to the increasing overlap between wavefunctions (and hence increasing scattering rate) as a consequence of stronger confinement.

4. Conclusions

In this paper, the self-consistent solution of Schrödinger and Poisson equations has been applied to the simulation of single- and double-gate SOI ultrathin MOS structures. The impact of quantization on the characteristics of these devices has been analyzed and the two structures have been compared to investigate the role played by volume-inversion as well as the reasons for the double-gate device's enhanced performance. Low field mobility calculations shed light on the non-trivial dependence of mobility on silicon layer thickness.

References

1. J.-P. Colinge, *Silicon on Insulator Technology: Materials to VLSI*, New York: Kluwer, 1990.
2. J.-H. Choi, Y.-J. Park, and H.-S. Min, "Electron mobility behavior in extremely thin SOI MOSFETs," *IEEE Electron Dev. Lett.* **16**, 527 (1995).

3. D. J. Frank, S. E. Laux, and M. V. Fischetti, "Monte Carlo simulation of a 30 nm dual-gate MOSFET: how short can Si go?" *IEDM Tech. Digest* (1992), p. 553.
4. F. Balestra, S. Cristoloveanu, M. Benachir, J. Brini, and T. Elewa, "Double-gate silicon-on-insulator transistor with volume inversion: a new device with greatly enhanced performance," *IEEE Electron Dev. Lett.* **8**, 410 (1987).
5. J.-P. Colinge, M. H. Gao, A. Romano-Rodriguez, H. Maes, and C. Claeys, "Silicon-on-insulator 'gate-all-around device'," *IEDM Tech. Digest* (1990), p. 595.
6. S. Venkatesan, G. W. Neudeck, and R. F. Pierret, "Dual-gate operation and volume inversion in n-channel SOI MOSFETs," *IEEE Electron Dev. Lett.* **13**, 44 (1992).
7. F. Balestra, "Comments on 'Dual-gate operation and volume inversion in n-channel SOI MOSFETs'," *IEEE Electron Dev. Lett.* **13**, 658 (1992).
8. S. Venkatesan, R. F. Pierret, and G. W. Neudeck, "Reply to comments on 'Dual-gate operation and volume inversion in n-channel SOI MOSFETs'," *IEEE Electron Dev. Lett.* **13**, 659 (1992).
9. Y. Omura, S. Nakashima, K. Izumi, and T. Izumi, "0.1 μm -gate, ultrathin-film CMOS devices using SIMOX substrate with 80-nm-thick buried oxide layer," *IEEE Trans. Electron Dev.* **40**, 1019 (1993).
10. A. Auberton-Herve, "SOI materials to systems," *IEDM Tech. Digest* (1996), p. 3.
11. A. Abramo, J. Bude, F. Venturi, and M. R. Pinto, "Mobility simulation of a novel Si/SiGe FET structure," *IEEE Electron Dev. Lett.* **17**, 59 (1996).
12. M. J. van Dort, P. H. Woerlee, A. J. Walker, C. A. H. Juffermans, and H. Lifka, "Influence of high substrate doping levels on the threshold voltage and the mobility of deep-submicrometer MOSFETs," *IEEE Trans. Electron Dev.* **39**, 932 (1992).
13. C. Jacoboni and L. Reggiani, "The Monte Carlo method for the solution of charge transport in semiconductors with applications to covalent materials," *Rev. Mod. Phys.* **55**, 645 (1983).
14. C. Jungemann, A. Edmunds, and W. L. Engl, "Simulation of linear and nonlinear electron transport in homogeneous silicon inversion layers," *Solid State Electron.* **32**, 1529, (1993).

Can Silicon-Based Heterodevices Compete with CMOS for System Solutions?

E. Kasper and G. Reitemann

Institut für Halbleitertechnik, Universität Stuttgart, Pfaffenwaldring 47, D-70569 Stuttgart, Germany

1. Introduction

In a rapidly developing field like microelectronics, both researchers at the cutting edge and manufacturers who develop commercial products feel the need for general rules that help to extrapolate from the past into the future. The most prominent rule is known as Moore's Law, which states that on a semilogarithmic scale the critical device dimensions shrink and the circuit complexity increases linearly with time. Complexity in this context is represented by the number of transistors in certain integrated circuits (ICs), like memories or microprocessors. It should be noted, however, that initial exponential growth is not specific to microelectronics alone, but applies to many emerging technologies. Still, one marvels at the steep slope and long duration (over 30 years) of the exponential growth, which has been accompanied by continuous price reduction per transistor in ICs. Eventually, rapid evolution of IC complexity will be enhanced by integrating different components, leading to systems on a chip.

System integration requires rethinking of the integration scheme beyond the current model of component integration, in which CMOS technology provides highly perfect logic circuits while input/output circuits lag behind. In this article we will consider systems with components that differ significantly either in their electrical functions (e.g. analog/digital) or in their nonelectrical functions (e.g. mechanical/optical) or in their power/voltage level (power/logic) or in their technology (e.g. bipolar/CMOS).

Key issues for system solutions concern the high system complexity as well as economics. Will monolithic integration be the ultimate choice? Will CMOS prove an economically viable technology in areas other than digital computation? Will novel device structures allow unified solutions for different functions?

The paper is organized as follows. First we examine the advantages and drawbacks of different paths to system solutions. Then we explore the idea of a common device structure for different functions, using the particular example of a novel device that combines HBTs with charge injection transistors (CHINTs). Finally, we investigate the realisation possibilities of such devices using Si/SiGe heterostructures.

2. Different paths to system solutions

Today, when you look at a typical system you will find many ICs, discrete devices, and passive elements mounted and interconnected on a board. We will not discuss this conventional solution, except to note that packing density, interconnects, and mounting techniques are improving continuously. Integration can be achieved in hybrid multichip modules (MCMs) or monolithically.

3. Multi-chip modules

Different materials (laminates, ceramics, silicon) can be used as substrates, where the chips are mounted and interconnected. Very elegant solutions were obtained with silicon as substrate, since the thermal expansions of chip and substrate are the same and lithography and metallization for interconnects follows the same principles as in chip fabrication. A wide variety of different chips have already been integrated. As an example, Fig. 1 shows the integration of an active antenna for a mm-wave receiver (96 GHz) with a low-frequency CMOS circuit.¹ The multi-chip module technique has already obtained a good market acceptance and will be a serious candidate for system integration. Advantages are the separate optimization of chips and the sourcing of chips from different manufacturers; drawbacks are limited availability and testing of unpackaged chips, the interconnect length, and the additional mounting effort.

4. Monolithic integration

Monolithic integration of systems is probably restricted to silicon substrates, because of the importance of high complexity logic only available in silicon. A possible list of technical solutions includes:

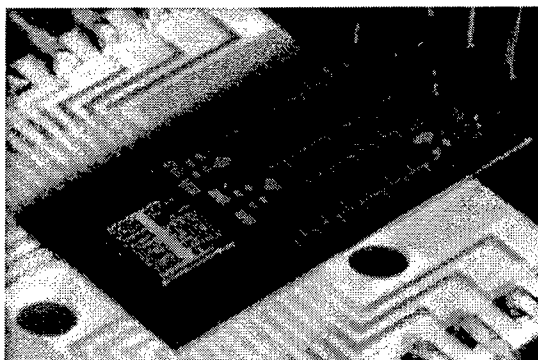


Figure 1. Multichip module with an active mm-wave rectifying antenna (rectenna) and a CMOS amplifier mounted on a high resistivity silicon substrate.¹

- Chips entirely made in CMOS
- Bipolar and CMOS (BiCMOS)
- Subchip attachment techniques
- Heterointegration

Logic circuits are dominated by CMOS transistors, because in this technology the basic unit – the inverter – consumes very small power at static operation. In a simple theory, assuming field-independent mobility and operation in the saturation regime, the current I_D is given by

$$I_D = \beta(V_{GS} - V_{TH})^2 \quad (1)$$

where $\beta = (\mu W \epsilon_{ox}) / (2L d_{ox})$, and μ is the mobility, W is the device width, L is the gate length, ϵ_{ox} and d_{ox} are the oxide permittivity and thickness, V_{GS} and V_{TH} are the gate-source and threshold voltages. For the cut-off frequency f_T one then finds

$$2\pi f_T = \mu(V_{GS} - V_{TH}) / L^2 = g_m / C_{ox} \quad (2)$$

where g_m is the transconductance and C_{ox} the gate oxide capacitance. The gate length shrinkage leads to smaller footprint, higher transconductance, and higher frequency limits. Today, production of microprocessors utilizes CMOS with 0.25 μm gate length.² Projected downscaling of CMOS devices with the corresponding technical specifications and performance is described in roadmaps.³ The future $L = 0.1 \mu\text{m}$ CMOS transistor will be an impressive device with low voltage and respectable speed. Also, MOS transistors with high voltage and power handling capability can be built, but on different substrates and with different technology, making system integration entirely within CMOS a none-too-easy task.

As a result, a combination of bipolar and CMOS technology (BiCMOS) was suggested, realized and produced.⁴ Bipolar transistors can deliver large drive currents, operate with small logic swings, and have high noise immunity. A schematic diagram of a BiCMOS device is shown in Fig. 2.

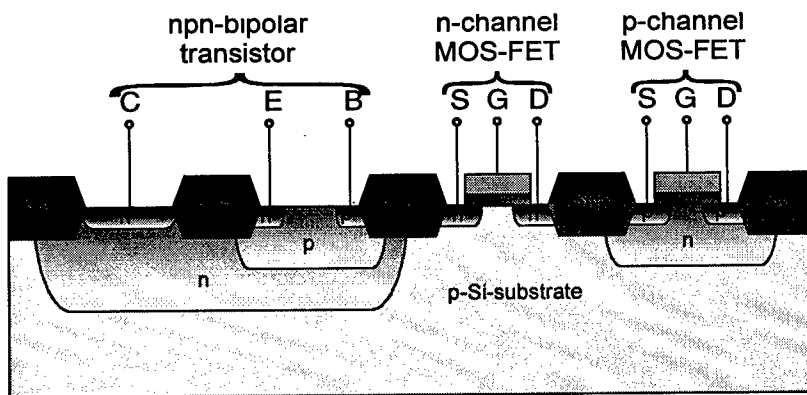


Figure 2. Schematic diagram of a BiCMOS circuit.

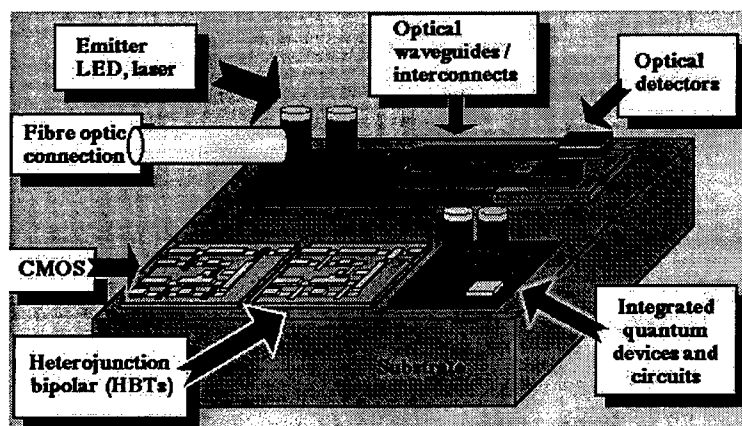


Figure 3. Sketch of a potential future chip with different integrated components.⁵

BiCMOS offers the benefits of both bipolar and CMOS circuits by appropriately trading off the characteristics of each technology, but at the cost of added process complexity. If the system were to integrate different semiconductor technologies in addition to silicon, the process complexity would increase further. A sketch of a future chip⁵ integrating many different devices and circuits would be fascinating indeed — see Fig. 3.

However, the limited success of BiCMOS should be a serious warning that adding process complexity will not easily be accepted. The devices from different materials could be added by subchip attachment techniques or directly fabricated by epitaxial heterostructure growth on top of the silicon device level (heterointegration). A possible scheme for heterointegration would consist of three main steps.⁶ In the first step, the silicon devices are fabricated without the metallization level and areas for the later heterodevices are defined, e.g. by oxide windows. In the second step, the heterostructures are grown within the windows and the heterodevices are fabricated with a small thermal budget process. In the final step, the common metallization and passivation is formed.

From the viewpoint of process complexity — and BiCMOS is teaching us how important this topic is — a common structure for different devices would be a highly desirable target. Obviously this target cannot be realized with a single material system, however let us think about possible solutions with heterostructures.

5. The search for a common device structure

Heterostructures offer numerous possibilities for enhancing device performance and progress is ongoing in many areas. The idea behind this section is to point to the unique potential of heterostructures for a unified system technology. This direction is a new one for heterostructure research, with system performance and

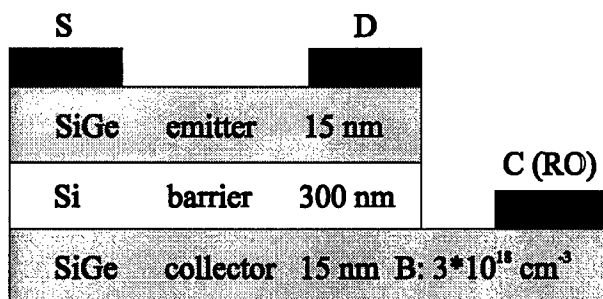


Figure 4. Diagram of a *p*-type SiGe charge injection transistor (CHINT).⁷

low process complexity as the primary goals. The idea is explained with a specific example, but we hope to stimulate other suggestions.

Let us combine as a primary building block a field effect transistor or a real space transfer device and a bipolar transistor. We have chosen a charge injection transistor (CHINT) as real space transfer device and an HBT for the bipolar side. A schematic diagram of a CHINT⁷ is presented in Fig. 4.

On top of a Si substrate there is essentially a three layer structure SiGe/Si/SiGe called collector/barrier/emitter. The source/drain contacts are on the top emitter, the collector contact is on the bottom collector layer. By real space transfer hot carriers accelerated by the drain-source voltage cross the barrier to the collector when an appropriate voltage is applied. In the following we denote the CHINT collector contact with RO (real space transfer output) to avoid confusion with the bipolar collector contact. In a common-source configuration of the CHINT the usual input is the drain and RO is the output, therefore the drain in the CHINT is more comparable to the gate electrode in a field effect transistor. A variety of logic functions can be realised with CHINT transistors.⁷ The HBT layer structure on the silicon substrate⁸ consists also of a three layer structure, however in the order Si/SiGe/Si. The proposed common layer structure is shown in Fig. 5. It consists of a four layer *n*-Si/*p*⁺-SiGe/Si/SiGe structure on a Si substrate with an *n*⁺ buried layer beneath the bipolar area. The *p*⁺-SiGe layer is the base of the HBT and the collector of the CHINT. As shown in Fig. 5, the transistor areas can be defined by trench etching and the base (B), emitter (E) and RO contacts are performed by implantation or diffusion from a polysilicon source, while the HBT collector is contacted via a buried layer subcollector (C).

6. Silicon based heterostructures

Amongst several silicon-based heterostructures (e.g. silicides NiSi and CoSi₂, insulators like CaF₂, semiconductors like GaP or SiC), the silicon-germanium on

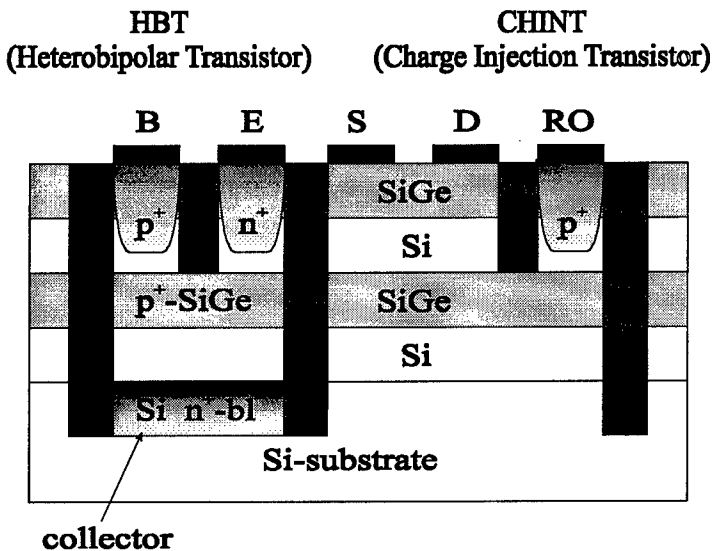


Figure 5. Common layer structure for an integrated HBT and CHINT on a Si substrate. HBT contacts are on the emitter (E), base (B) and on the collector outside of the plane via the buried layer subcollector. CHINT contacts are on the source (S), drain (D) and on the CHINT collector (RO). This contact is denoted RO to distinguish it clearly from the HBT collector contact.

silicon (SiGe/Si) turned out to be most practical because of its chemical similarity, its complete miscibility and its moderate lattice mismatch.⁹ As a result, SiGe/Si heterostructures are compatible with Si technology. These heterostructures act as carrier filters in single-barrier structures, exploit quantum effects in double-barrier structures, extend the optical sensitivity into the infrared, and promise fabrication of sub-100 nm structures by self-ordering. The obstacles to widespread technological adoption are the relaxation of layers above a critical thickness h_c and reduced thermal budget processing. Layers below the equilibrium critical thickness are elastically strained and stable under all process conditions. With growth at reduced temperatures (e.g. 550 °C) and processing at reduced temperatures (e.g. 750–850 °C for Si capped structures), significantly higher critical thickness can be maintained (metastable regime). For instance, for a 25% Ge content alloy on Si the equilibrium critical thickness amounts to 6 nm, whereas metastable growth at 550 °C is possible up to 70 nm.⁹

Recent device research into SiGe has focused on SiGe HBTs for mobile communication applications. High frequency limits ($f_T, f_{max} > 100$ GHz) and low noise levels were obtained by making use of the high current drive and very narrow highly-doped base layers possible in HBTs. The next attractive device target would be a hetero-CMOS with high symmetrical mobilities (1500–2500

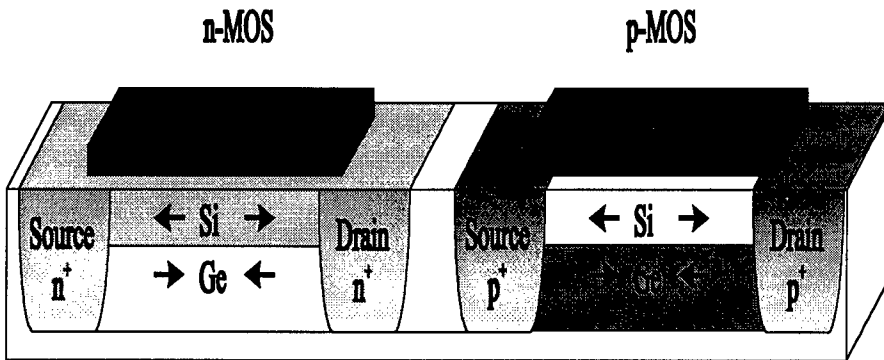


Figure 6. Scheme of a hetero-CMOS device. The strain in the layers is indicated by arrows (Si under tensile strain, Ge under compressive strain). The *n*-MOS (left) has a electron channel in Si, the *p*-MOS (right) has a hole channel in Ge.

$\text{cm}^2/\text{V}\cdot\text{s}$) for holes and electrons.¹⁰ In Si CMOS the mobilities are moderately high but asymmetrical, with electron $\mu \approx 800 \text{ cm}^2/\text{Vs}$ and hole $\mu \approx 250 \text{ cm}^2/\text{Vs}$. Elastically strained SiGe is the only material system that offers the desirable property of high symmetrical mobilities for both carriers (holes in strained Ge, electrons in strained Si). Figure 6 shows a schematic layout of a hetero-CMOS device.

In strained Si/Ge (with Ge under compressive and Si under tensile strain) a type II heterointerface lets the electrons jump in the Si channel and the holes in the Ge channel, respectively.¹⁰ The strain status is obtained by a virtual substrate and relaxed SiGe layer on top of it. The realization of this high performance hetero-CMOS requires research on improvement of the crystal quality of the virtual substrate and on growth of SiGe layers with high Ge content.

7. Outlook

Within a few years the SiGe-HBT has emerged as the fastest Si-based transistor and several companies are now prepared to produce HBT based ICs. A scheme for a superior hetero-CMOS is available. Strain adjustment by virtual substrates and growth of highly strained layers need additional research efforts. Silicon-germanium based optoelectronics will also benefit from research in this area. System integration with demand for different devices will be of increasing importance. A unified heterostructure technology with a common layer structure for different devices could solve obvious problems for system integration and may compete successfully with CMOS in this growing segment of the semiconductor market.

References

1. K. M. Strohm, J. Buechler, and E. Kasper, "SIMMWIC rectennas on high-resistivity silicon and CMOS compatibility," *IEEE Trans. Microw. Theory* **46**, 669 (1998).
2. P. Singer, "The dawn of quarter micron production," *Semicond. International*, January 1997 issue, pp. 50-56 (1997).
3. Semiconductor Industry Association Roadmap, 1994.
4. C. Y. Chang and S. M. Sze, *ULSI Technology*, New York: McGraw-Hill, 1996.
5. D. J. Paul, "Silicon-germanium heterostructures in electronics: the present and the future," *Thin Solid Films* **321**, 172 (1998).
6. E. Kasper and J.-F. Luy, "Molecular beam epitaxy of silicon-based electronic structures," *Microelectronics J.* **22**, 5 (1991).
7. M. Mastrapasqua, C. A. King, P. R. Smith, and M. R. Pinto, "Functional devices based on real space transfer in Si/SiGe structures," *IEEE Trans. Electron Dev.* **43**, 1671 (1996).
8. E. Kasper, "Prospects of SiGe heterodevices," *J. Crystal Growth* **150**, 921 (1995).
9. E. Kasper, ed., *Properties of Strained and Relaxed Silicon Germanium*, EMIS Datareviews Series No. 12, London: IEE, INSPEC, 1995.
10. F. Schäffler, "High mobility Si and Ge structures," *Semicond. Sci. Technol.* **12**, 1515 (1997).

Si/SiGeC Heterostructures: A Path Towards High Mobility Channels

R. Hartmann, Ulf Gennser, H. Sigg, D. Grützmacher, and G. Dehlinger
Paul Scherrer Institute, CH-5232 Villigen-PSI, Switzerland

1. Introduction

Si/SiGe has shown itself a viable material system in production for bipolar technology. The question is whether Si/SiGe heterostructures can also be used to enhance Si MOSFETs. Both *p*-type and *n*-type quantum well channels have been demonstrated, with elevated low-temperature hole and electron mobilities respectively. Further, Si/SiGe *p*- and *n*-channel MODFETs have shown enhanced performance compared with Si MOSFETs.¹ However, there are still some large obstacles for the implementation of such devices in CMOS technology:

- In pseudomorphic structures grown on Si substrates, the Si/SiGe band offset lies mainly in the valence band, with a conduction band offset less than 10 meV for all reasonable Ge concentrations,² too small for any electron confinement. The conduction band offset is small because of coincidental cancellation of the intrinsic band offset by the effects of strain. In high mobility electron devices this cancellation has been avoided by growth on relaxed SiGe buffer layers,³ thereby changing the strain configuration. However, this carries with it new problems: the strain in the SiGe buffer layer relaxes through the formation of misfit dislocations. Though it is possible to grow buffer layers in a way that makes the dislocations penetrate preferentially down into the substrate, instead of through the SiGe layers, their density in the active layers is still too high to be acceptable for CMOS technology.
- The Si/SiGe material allows for a much smaller thermal budget than a normal fabrication sequence.⁴ This problem has been overcome in Si/SiGe HBTs by adjusting the design to be less sensitive to a smearing of the dopants. MODFETs are much more sensitive to the doping profile. Generally, the doping is placed 100 Å from the quantum well, and any doping diffusion into the well is detrimental to the performance.

In recent years, there have been large advances in the growth of pseudomorphic SiGeC alloys.⁵⁻⁷ The introduction of C into IV-IV heterostructures increases the design freedom, as C can reduce the compressive strain in the alloy layers, or even introduce a tensile strain. The strain reduction allows for much larger layers without exceeding the critical thickness. Since strain is a crucial factor for the band offsets, one can also expect that they will be largely affected by C alloying. It has also been found that the C atoms can reduce dopant diffusion significantly,⁸ due to a reduction of available interstitials. In this article we will argue that

SiGeC alloys may offer an increased leverage in the CMOS technology, just as SiGe has increased the performance of bipolar technology. We report on the use of photoluminescence spectroscopy to extract values for the band discontinuities of strain-reduced $\text{Si}_{1-x-y}\text{Ge}_x\text{C}_y$ MQWs. Through linear extrapolation these values are extended to the interesting region with $0 < x < 15\%$ and $0 < y < 1.5\%$. It is shown that both electron and hole confinement appear possible without the need of relaxed buffer layers, making the SiGeC alloy a potential for CMOS technology. Issues concerning the electronic material quality and the thermal processing budget are also discussed. Although it has so far been possible to introduce only very small fractions of C into these alloys, up to 1.5%, our results indicate that this may be enough to create a significant band offset in the conduction band, without relying on relaxed buffer layers. This effect, together with the diffusion quenching effects, makes SiGeC a very attractive topic for future investigations.

2. Experimental details

The Si/SiGeC MQWs discussed in this work have been grown by solid source molecular beam epitaxy (MBE), with electron guns for Si and Ge evaporation and a graphite filament for C sublimation. The growth was performed on Si(001) substrates at a temperature of 500 °C and a deposition rate of 1 Å/s for the Si and 0.3 Å/s for the SiGeC layers. The reduction in growth rate during the SiGeC deposition is necessary to allow C incorporation at sufficiently high concentrations.⁹ During growth a bias of +600 V applied to the substrate helps to reduce bombardment damage in the samples due to e-beam evaporation.¹⁰ Each sample consists of 6 periods with period lengths between 135 Å and 360 Å. The Ge concentration of the SiGeC layers is kept constant at 6% and the C content varies between 0% and 0.6% allowing different strain conditions in the layers. The structural quality of the layers is verified by transmission electron microscopy (TEM), and the composition is determined by high resolution x-ray diffraction assuming Vegard's law and linearly interpolated elastic constants. After the growth, the samples were treated by 10 minute anneals at 800 °C in forming gas. At these temperatures atomic diffusion and/or strain relaxation by defect formation or SiC precipitation do not occur.¹⁰ The photoluminescence (PL) is excited at 2.2K by 3 mW of Ar⁺ laser radiation ($\lambda = 488$ nm), dispersed by a grating monochromator and detected by a nitrogen cooled Ge photoconductor.

3. Analysis and discussion

Due to the differences in effective masses in the conduction band and the valence band, PL measurements on MQWs with constant compositions but different layer widths allow us to extract the band alignments. For this purpose a simple Kronig-Penney model is used. Since band bending due to photoexcitation is not observed in excitation dependent PL measurements, this effect is not considered in the calculations. Furthermore a rise of the sample temperature from 2 K to 12 K

does not affect the luminescence energy of the $\text{Si}_{0.936}\text{Ge}_{0.06}\text{C}_{0.004}$ signal; i.e. weak trapping of charge carriers by local strain fields in the layers seems not to occur. Therefore, within the range of C content studied here ($0 \leq y \leq 0.006$) local strain fields due to the small C radii are neglected. In addition, any Ge or C segregation during the deposition is neglected, and we assume a constant exciton binding energy of 15 meV.

For C concentrations of 0.4% and 0.6%, the widths of SiGeC (Si) layer has been varied between 27 Å and 180 Å, while keeping the Si (SiGeC) layer width constant. Keeping the SiGeC layer thickness constant at 45 Å and changing the Si layer width does not affect the no-phonon line. Instead, the luminescence signal shifts monotonically up in energy when the $\text{Si}_{0.936}\text{Ge}_{0.06}\text{C}_{0.004}$ width is reduced from 180 Å to 27 Å (Fig. 1). This behaviour shows that the PL signals originate from quantum confined subband levels with a type-I band alignment, where the ternary layer is acting as the quantum well.

The values for the band offsets are deduced from the experimental energy shifts by model calculations. For the $\text{Si}_{0.936}\text{Ge}_{0.06}\text{C}_{0.004}$ quantum well, the effective $\Delta(4)$ electron mass of Si bulk material $m_{\Delta(4)} = 0.19m_0$ and the effective SiGe_{0.06} heavy hole mass $m_{hh} = 0.27m_0$ are used taking the band edge splitting due to strain into account. For the barrier we take the effective mass of the transverse $\Delta(2)$ Si

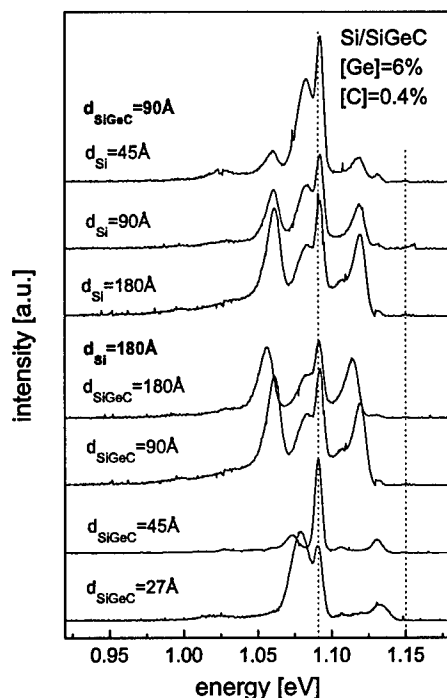


Figure 1. Low-temperature (2.2 K) photoluminescence spectra of 6 period $\text{Si}/\text{Si}_{0.936}\text{Ge}_{0.06}\text{C}_{0.004}$ MQW structures with different periods. The fixed positions of the Si related signals are indicated by the dotted lines. Typical type-I band alignment is observed for the strain reduced $\text{Si}_{0.936}\text{Ge}_{0.06}\text{C}_{0.004}$ structure.

electron state $m_{\Delta(2)} = 0.92m_0$. The use of the $\Delta(2)$ mass seems reasonable, since one can expect significant interface scattering of the charge carriers, as well as alloy scattering due to the large local strain in the vicinity of the C atoms. Furthermore, fitting the data assuming a conservation of the $\Delta(4)$ nature of the electrons in the barrier does not give physically reasonable results. Assuming type-I band character, the Kronig-Penney model yields the best fit for the $\text{Si}_{0.936}\text{Ge}_{0.06}\text{C}_{0.004}$ PL data for $\Delta E_{\text{CB}} = 21.1$ meV and $\Delta E_{\text{VB}} = 21.5$ meV (Fig. 2). For the samples with 0.6% C content, the behaviour of the PL energy is described best by assuming type-I band alignment with band offsets $\Delta E_{\text{CB}} = 33.2$ meV in the conduction band and $\Delta E_{\text{VB}} = 9.0$ meV in the valence band.

Since band offsets and band gap independently vary with Ge content, C content and strain in the layers, the knowledge obtained from the optical characteristics is limited to the two specific compositions used in the experiments, with $x = 0.06$ and $y = 0.004$ or $y = 0.006$, respectively. We will now develop a more comprehensive picture of the band offsets covering the whole range of Ge concentration x between 0% and 15% and C concentration y between 0% and 1.5%, by interpolating between and extrapolating from these values. While keeping in mind the danger of extrapolations, we can construct a band offset map. An encouraging sign of the map's validity is the linearity of the bandgap dependence on C and Ge content. Also, the extrapolation of the band offsets to different C contents agrees with the PL data on pseudomorphic 45 Å $\text{SiGe}_{0.06}\text{C}_y$ MQWs¹¹ for different C contents up to 0.8% and explains the bandgap reduction of strain-compensated SiGeC compared to compressive SiGeC.^{11,12}

The interaction of the light hole states with spin orbit split holes makes the valence band splitting non-linear, and requires the extrapolation of the band offsets to be done for unstrained materials. Model-solid theory¹³ can be used to calculate

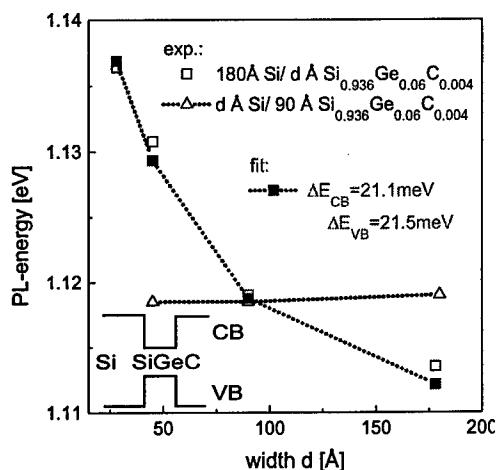


Figure 2. Dependence of the $\text{Si}_{0.936}\text{Ge}_{0.06}\text{C}_{0.004}$ NP energy as a function of d_{SiGeC} and d_{Si} , respectively. A type-I band alignment with SiGeC acting as the quantum well (inset) and band offsets $\Delta E_{\text{CB}} = 21.1$ meV and $\Delta E_{\text{VB}} = 21.5$ meV fits the experimental data the best.

the strain-induced band offset components in $\text{Si}/\text{Si}_{0.936}\text{Ge}_{0.06}\text{C}_{0.004}$ and $\text{Si}/\text{Si}_{0.934}\text{Ge}_{0.06}\text{C}_{0.006}$. In Fig. 3 the remaining, unstrained discontinuities are shown as a function of the C content. The offsets between unstrained Si and $\text{Si}_{0.94}\text{Ge}_{0.06}$ are calculated from data of Refs. 2, 14. The intrinsic offsets show a remarkably linear dependence on the C content, and the linear fits can be used to calculate the offsets for $0 < y < 0.015$. Using the same method as for the $\text{SiGe}_{0.06}\text{C}_y$ data in Fig. 1, the band offsets between unstrained Si and $\text{Si}_{1-y}\text{C}_y$ bulk materials can be taken from PL measurements on $\text{Si}/\text{Si}_{0.99}\text{C}_{0.01}$ MQW structures (where Si layers were constant at 156 Å, while the $\text{Si}_{0.99}\text{C}_{0.01}$ layers varied between 11 and 110 Å).¹⁵ The $\text{Si}/\text{Si}_{1-y}\text{C}_y$ offsets are included in Fig. 3 and show a similar behaviour as $\text{SiGe}_{0.06}\text{C}_y$. The incorporation of C evidently lowers the conduction band and valence band edge energies and decreases the intrinsic bandgaps in unstrained $\text{Si}_{1-y}\text{C}_y$ and $\text{SiGe}_{0.06}\text{C}_y$.

Figure 3 indicates that the intrinsic $\text{Si}/\text{Si}_{1-y}\text{C}_y$ band offsets depend linearly on the C content. We now make the additional assumption, that they also have a linear dependence on the Ge content. It is then possible to linearly interpolate between the two lines for $\text{Si}_{1-y}\text{C}_y$ and $\text{SiGe}_{0.06}\text{C}_y$, for the alloys with $0 < x < 0.06$ and $0 < y < 0.008$. Through a linear extrapolation, this is extended to the experimentally relevant region $0 < x < 0.15$ and $0 < y < 0.015$. Although this linear approximation seems reasonable for a certain range of Ge and C content, it is clear that the scant experimental data warrants future investigations. In Fig. 4 (conduction band) and Fig. 5 (valence band) the offsets are presented, when the strain effects are re-incorporated. The band offsets are shown as contour lines illustrating the layer compositions with constant band discontinuities. The kinks in the equi-offset lines in Fig. 4 correspond to the strain compensated alloys, where the $\Delta(2)$ and $\Delta(4)$ conduction bands are degenerate. To the left or right of

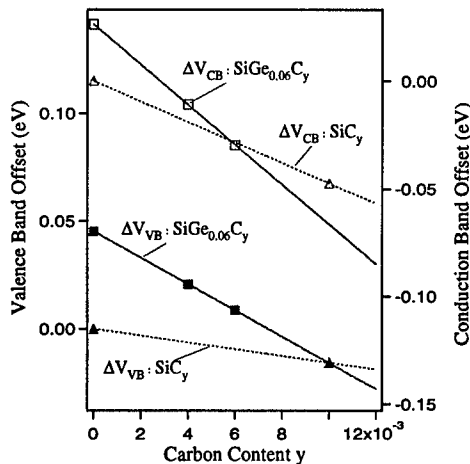


Figure 3. Band offsets for unstrained $\text{Si}_{0.936}\text{Ge}_{0.06}\text{C}_y$ and SiC_y as a function of C content y . The offsets are given relative to the Si band edges. The band offsets for $\text{Si}_{0.936}\text{Ge}_{0.06}\text{C}_y$ with $y = 0.004$ and $y = 0.006$ are taken from our PL measurements, whereas the values for $\text{Si}_{0.94}\text{Ge}_{0.06}$ and $\text{Si}_{1-y}\text{C}_y$ are taken from literature.^{2,14,15}

these kinks the alloys are either tensilely or compressively strained with the band minima constituted by the $\Delta(2)$ and $\Delta(4)$ band, respectively.

Let us first consider how the offsets move with the Ge concentration, for a constant C content, starting from the position of the strain compensated kink. For a reduction in x , there will be an increase in the tensile strain, pulling down the $\Delta(2)$ conduction band and lifting up the light hole valence band. Simultaneously, the unstrained, degenerate SiGeC conduction band moves up with respect to the Si band for C content $y < 0.006$, and decreases for $y > 0.006$ (moving vertically in Fig. 3). Thus, for $y < 0.006$, where the strain and intrinsic effects push the offset in the same direction, a reduction of the Ge concentration in strain-compensated SiGeC increases the conduction band offset (Fig. 4). With increasing y ($0 < y < 0.006$) the intrinsic impact on the band energy decreases and completely vanishes for $y = 0.006$. In this case the strain alone determines the variation of the conduction band offset of the $\text{Si}/\text{Si}_{1-x-y}\text{Ge}_x\text{C}_y$ heterostructure, where the $\text{Si}_{1-x-y}\text{Ge}_x\text{C}_y$ is under tensile strain. For higher C content, the intrinsic and strain induced effects will start to cancel each other, until, for $y > 0.009$ the intrinsic shifts become dominant, and the conduction band offset decreases with decreasing Ge content x . In the valence band (Fig. 5), the intrinsic and stress-induced effects are balancing each other out, except for quite high Ge concentrations, and for most alloys with tensile strain the valence band offset is not very large.

If we now instead increase the Ge content, the compressive strain will shift down the $\Delta(4)$ conduction band and lift up the heavy hole valence band with respect to the Si band edge. Intrinsic conduction band offset and strain effects are moving in the same direction for $y > 0.006$. Thus, the conduction band offset strongly increases with the Ge content x . For $y < 0.006$, the addition of Ge shifts the unstrained conduction band edge to higher energy and thus reduces the strain-

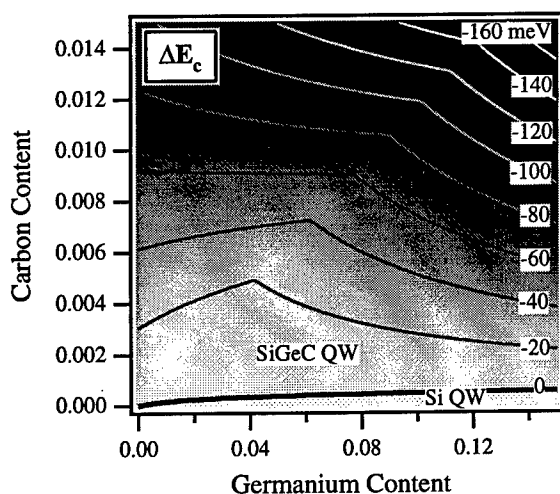


Figure 4. Energy contour plots of the $\text{Si}_{1-x-y}\text{Ge}_x\text{C}_y$ conduction band edge relative to the Si band edge as a function of Ge and C concentration.

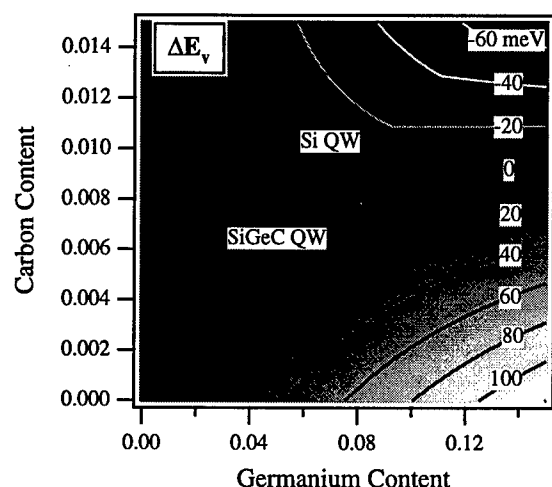


Figure 5. Energy contour plots of the $\text{Si}_{1-x-y}\text{Ge}_x\text{C}_y$ valence band edge relative to the Si band edge as a function of Ge and C concentration.

induced energy decrease of the $\Delta(4)$ state. This effect is seen in Fig. 4, where the slope of the contour lines in the compressive strain region decreases with decreasing C content in SiGeC. However, whereas for the tensile region a compensation between the intrinsic and strain induced effects is reached for $y = 0.009$, this balance is not achieved for compressive strain until $y \approx 0$. This is, of course, the well-known "negligible" conduction band offset for Si/SiGe heterojunctions. The valence band offset in the compressive region is almost solely determined by the intrinsic band offset, the uniaxial strain and hydrostatic pressure shifts of the heavy hole band acting in opposing directions. It is observed that, for sufficiently high C content, the valence band offset shows a confinement in the Si layer instead of the SiGeC layer. This effect can already be surmised from the fits of the PL from pseudomorphic Si/SiGe_{0.06}C_y MQWs, where a decreasing valence band offset is found for increasing C content (Fig. 3). Whether and where this leads to an inversion of the Si and SiGeC valence band alignment depends on the details of the extrapolation, and needs to be investigated in further experiments.

The variation in the conduction band alignment may seem surprising: SiGe is known to have a very small conduction band offset, and when increasing the Ge content in the $\text{Si}_{1-x-y}\text{Ge}_x\text{C}_y$ alloy one would think that the offsets would be closer to that of the $\text{Si}_{1-x}\text{Ge}_x$ alloy. However, as we have seen, the addition of C suppresses the balance between the intrinsic offset and the strain effects. Consequently, care has to be taken to separate different effects when using the simple picture of ternary SiGeC as an intermediate between SiGe and SiC.

As we have stated above, the Figs. 4 and 5 have to be taken with a grain of salt, because of the extrapolations made. However, they indicate the possibility of obtaining respectably large conduction band offset: either in a Si/SiC structure

with sufficiently high C content or by increasing the Ge content in the alloys. Until further measurements, the values of the conduction band offsets are still speculative, but as long as the offset retains a linear dependence on the Ge and C contents, our analysis indicates that if the offset in an alloy with higher Ge content is smaller than given here, the offset in the SiC alloy will be larger, and *vice-versa*. A disadvantage of the Ge-rich alloy is that the conduction band will be determined by the $\Delta(4)$ bands, which have a rather heavy effective mass in the plane of the interfaces. We are currently investigating the low temperature transport characteristics of *n*-type modulation-doped SiC and SiGeC structures.

4. Possibilities for MOSFET technology

We have seen that there are two possibilities for electron confinement in SiGeC heterostructures on Si substrates: either in tensilely strained SiC quantum wells with sufficiently high C-content, or in compressively strained SiGeC quantum wells. In addition, it has been shown that hole confinement is possible in strain-reduced SiGeC channels. However, this article would be rather incomplete without a discussion about the feasibility of using SiGeC channels in Si MOSFET technology. Two issues are especially relevant: can SiGeC quantum wells be grown with sufficiently high electrical quality, and does the material allow for a high enough thermal budget?

There is a concern that the large lattice mismatch between Si and C may create local distortions of the lattice, as well as point defects, leading to a severe reduction of the mobility. Initial experiments indicate that sufficiently good quality material may be obtained. Faschinger and co-workers have grown $\text{Si}_{0.988}\text{C}_{0.012}$ layers with an *n*-type background doping of $4 \times 10^{16} \text{ cm}^{-3}$ that show mobilities comparable to that of bulk Si with similar doping levels.¹⁶ On the other hand, mobility measurements on *n*-type SiC and *p*-type SiGeC layers grown at 400 °C have shown a decreasing mobility with increasing C content. This decrease has been attributed to a large fraction of C atoms forming neutral or negatively charged interstitials during the rather low temperature growth.¹⁷ Photoluminescence from SiC and SiGeC alloys is also weakened by the formation of C point defects, but it has been shown that their density is reduced by a 10 minute post-growth anneal at 800 °C,⁹ a procedure that we also used in the present experiments. Modulation-doped Si/SiGeC two-dimensional hole gases (2DHGs) show an increase in their mobility after similar anneals. In comparing the mobilities between the SiGe and the SiGeC 2DHGs, one should also bear in mind that the lower strain in the latter leads to a smaller confinement, and possibly also to a larger interband scattering between heavy and light holes. Therefore, there are grounds for optimism that SiGeC layers with adequate material quality for device applications are possible.

The thermal budget for *n*-channel Si/SiGe MODFETs has been investigated by König and co-workers, who found that rapid thermal anneals of 10 s should not exceed 600 °C.⁴ The dominant problem with high temperature processing of SiGe modulation doped QWs is believed to be dopant diffusion into the well. The

most commonly used p -dopant is B, which diffuses primarily through an interstitial mechanism. It has been found that the addition of C in a B-doped Si layer reduces the diffusion considerably. The reason for this is believed to be due to the formation of C interstitials, which reduces the possible diffusive sites for the B atoms.⁸ Apparently, in these experiments, the reduction in the number of point defects due to the anneal is not sufficient to destroy the diffusion quenching mechanism. The prevailing n -type dopant for MBE material, Sb, diffuses through mechanisms similar to those of B. However, an increase in interstitials can decrease the diffusion through the Frenkel reduction of vacancies, in contrast to the case of B diffusion. In extrinsic Si, the charged vacancies and Si interstitials play the same role for B doping and Sb doping. To our knowledge, Sb-doped layers have not been investigated. The results indicate possibilities at least for p -type Si/SiGe and Si/SiGeC MODFET, for complementary logic gates.

4. Conclusions

SiC and SiGeC alloys offer additional freedom for bandgap engineering. We have proposed a "map" for the band offsets of $\text{Si}/\text{Si}_{1-x-y}\text{Ge}_x\text{C}_y$. Our preliminary results indicate the possibilities of both p -type and n -type confinement. Though the demonstration of epitaxial SiGeC alloys is relatively recent, the material quality obtained today leaves room for optimism. The introduction of C also offers advantages for processing. We believe that, as a complementary alternative to the Si MOSFET roadmap, SiGeC MODFETs is a path well worth exploring.

References

1. M. Arafa, P. Fay, K. Ismael, J. O. Chu, B. S. Meyerson, and I. Adesida, "High speed p -type SiGe modulation doped field effect transistors," *IEEE Electron Dev. Lett.* **17**, 124, (1996);
U. König, A. J. Boers, F. Schäffler, and E. Kasper, "Enhancement mode n -channel Si/SiGe MODFET with high intrinsic transconductance," *Electronics Lett.* **28**, 160 (1992).
2. D. J. Robbins, L. T. Canham, S. J. Barnett, A. D. Pitt, and P. Calcott, "Near-band-gap photoluminescence from pseudomorphic $\text{Si}_{1-x}\text{Ge}_x$ single layers on silicon," *J. Appl. Phys.* **71**, 1407 (1992).
3. E. A. Fitzgerald, Y. H. Xie, M. L. Green, *et al.*, "Totally relaxed $\text{Ge}_x\text{Si}_{1-x}$ layers with low threading dislocation densities grown on Si substrates," *Appl. Phys. Lett.* **59**, 811 (1991);
F. K. LeGoues, B. S. Meyerson, and J. F. Morar, "Anomalous strain relaxation in SiGe thin films and superlattices," *Phys. Rev. Lett.* **66**, 2903 (1991).
4. U. König, A. J. Boers, and F. Schäffler, "N-channel Si/SiGe MODFET's: effects of rapid thermal activation on the dc performance," *IEEE Electron Dev. Lett.* **14**, 97 (1993).

5. K. Eberl, S. S. Iyer, S. Zollner, J. C. Tsang, and F. K. LeGoues, "Growth and strain compensation effects in the ternary $\text{Si}_{1-x-y}\text{Ge}_x\text{C}_y$ alloy system," *Appl. Phys. Lett.* **60**, 3033 (1992).
6. P. Boucaud, C. Guedji, D. Bouchier, *et al.*, "Optical properties of bulk and multi-quantum well SiGeC heterostructures," *J. Cryst. Growth* **157**, 410 (1995).
7. K. Eberl, S. S. Iyer, and F. K. LeGoues, "Strain symmetrization effects in pseudomorphic $\text{Si}_{1-y}\text{C}_y/\text{Si}_{1-x}\text{Ge}_x$ superlattices," *Appl. Phys. Lett.* **64**, 739 (1994).
8. L. D. Lanzerotti, J. C. Sturm, E. Stach, R. Hull, T. Buyuklimanli, and C. Magee, "Suppression of boron outdiffusion in SiGe HBTs by carbon incorporation," *Tech. Dig. IEDM*, 249 (1996);
H.J. Osten, B. Heinemann, D. Knoll, G. Lippert, and H. Rücker, "Effects of carbon on boron diffusion in SiGe: Principles and impact on bipolar devices," *J. Vac. Sci. Technol. B* **16**, 1750 (1998).
9. R. Hartmann, D. Grützmacher, E. Müller, *et al.*, "Growth of $\text{Si}_{1-y}\text{C}_y/\text{Si}$ - and $\text{Si}_{1-x-y}\text{Ge}_x\text{C}_y/\text{Si}$ - multiple quantum wells using molecular beam epitaxy," *Thin Solid Films* **318**, 158 (1998).
10. R. Hartmann, D. Grützmacher, E. Müller, U. Gennser, and A. Dommann, "Effects of substrate bias and rapid thermal processing on the luminescence of Si/SiGe multiple quantum wells grown by MBE," *Thin Solid Films* **294**, 50 (1997);
D. Grützmacher, R. Hartmann, P. Schnappauf, *et al.*, "Low temperature molecular beam epitaxy of $\text{Si}_{1-x-y}\text{Ge}_x\text{C}_y/\text{Si}$ - quantum well structures: electrical and optical properties," to be published in *Thin Solid Films* (1998).
11. R. Hartmann, U. Gennser, H. Sigg, D. Grützmacher, and K. Ensslin, "Band gap and band alignment of strain reduced Si/Si $_{1-x-y}\text{Ge}_x\text{C}_y$ multiple quantum well structures obtained by photoluminescence measurements," to be published in *Appl. Phys. Lett.* (1998).
12. K. Eberl, K. Brunner, and W. Winter, "Pseudomorphic $\text{Si}_{1-y}\text{C}_y$ and $\text{Si}_{1-x-y}\text{Ge}_x\text{C}_y$ alloy layers on Si," *Thin Solid Films* **294**, 98 (1997);
O. G. Schmidt and K. Eberl, "Photoluminescence of tensile strained, exactly strain compensated and compressively strained $\text{Si}_{1-x-y}\text{Ge}_x\text{C}_y$ layers on Si," *Phys. Rev. Lett.* **80**, 3396 (1998).
13. C. G. Van de Walle, "Band lineups and deformation potentials in the model-solid theory," *Phys. Rev. B* **39**, 1871 (1989).
14. M. L. W. Thewalt, D. A. Harrison, C. F. Reinhart, J. A. Wolk, and H. Lafontaine, "Type II band alignment in $\text{Si}_{1-x}\text{Ge}_x/\text{Si}(001)$ quantum wells: the ubiquitous type I luminescence results from band bending," *Phys. Rev. Lett.* **79**, 269 (1997).
15. K. Brunner, K. Eberl, and W. Winter, "Near-band-edge photo-luminescence from pseudomorphic $\text{Si}_{1-y}\text{C}_y/\text{Si}$ quantum well structures," *Phys. Rev. Lett.* **76**, 303 (1996).
16. W. Faschinger, S. Zerlauth, G. Bauer, and L. Palmetshofer, "Electrical properties of $\text{Si}_{1-x}\text{C}_x$ alloys and modulation doped $\text{Si}/\text{Si}_{1-x}\text{C}_x/\text{Si}$ structures," *Appl. Phys. Lett.* **67**, 3933 (1995).
17. H.J. Osten and P. Gaworzewski, "Charge transport in strained $\text{Si}_{1-x}\text{C}_x$ and $\text{Si}_{1-x-y}\text{Ge}_x\text{C}_y$ alloys on Si(001)," *J. Appl. Phys.* **82**, 4977 (1997).

Potential of SiGe-Channel MOSFETs for a Submicron CMOS Technology

J. Alieu*, T. Skotnicki, P. Bouillon, J. L. Regolini

France-Telecom, CNET Grenoble, B. P. 98, 38243 Meylan Cedex France

**Presently with Centre CNET - ST Microelectronics, 850 rue J. Monet, 38926 Crolles Cedex France and INSA Lyon*

A. Souifi, G. Guillot, and G. Brémont

INSA Lyon, LPM, UMR CNRS 511, Bat. 502, 69621 Villeurbanne Cedex France

1. Introduction

In order to achieve a big improvement in transistor performance and to increase the density of integration, silicon technology needs some breakthroughs with respect to advanced technology development. Concerning performance, the use of strained layers has been proposed to greatly improve carrier mobility.¹ The use of SiGe is the easiest way to integrate strained layers within the Si technology. Now it is possible to realize devices with a band gap engineering design without much change with respect to the standard CMOS process.

Several studies have been done on PMOS transistors only, fabricated with a low-temperature process (thick deposited gate oxide, low temperature anneal for dopant activation). The Ge concentration as well as the Ge profile have been studied in depth, always with a modified process. All these works have clearly demonstrated that very good results can be obtained with such low thermal budget processes.²⁻⁵ However, we have to consider the use of a SiGe channel architecture within a 0.12/0.18 μm complete CMOS process.

This paper reports on SiGe channel integration within a standard full CMOS process. After the structure description, we describe our process and then present our $C(V)$, $G(V)$, and DLTS measurements that characterize the layer quality. Next we show the electrical characteristics of PMOS long-channel transistors as a function of temperature, including our finding of a strong improvement in the hole mobility (190% at $T = 300\text{ K}$). These results are used to extract the valence band offset and the scattering parameters by simple and reliable methods. Finally, we present results on short-channel transistor operation that reveal the two limiting factors for SiGe integration within a standard full CMOS Si technology. In the conclusion, we propose an optimal architecture for the future generations of transistors that integrate such strained SiGe channel structures.

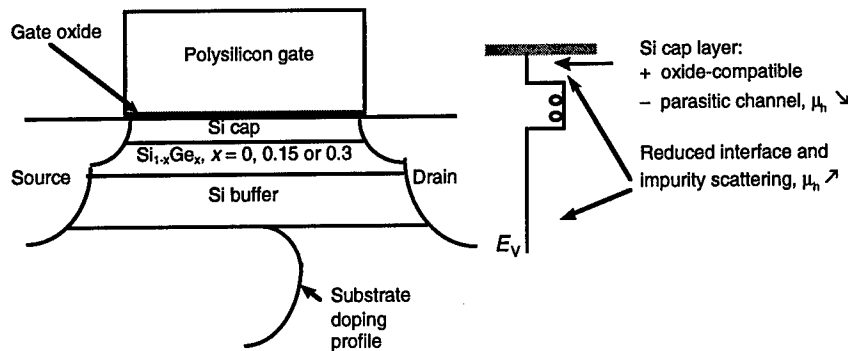


Figure 1. Epitaxial layer sequence and device layout for the SiGe transistor (left) and the corresponding valence band diagram in the channel (right).

2. Transistor design

Earlier studies of hole mobility improvement in SiGe material attributed the effect to the lowering of the longitudinal effective mass for holes. Our transistor active structure, shown schematically in Fig. 1, was designed to take advantage of high carrier mobilities. The three epitaxial layers were not intentionally doped and their thicknesses were set by the dopant diffusion for the Si buffer, the critical thickness function of the Ge fraction for the SiGe channel, and by the oxidation and Ge diffusion for the Si cap. The reasoning behind these technological parameters ran as follows:

- the buffer layer reduces the scattering on the impurities previously implanted into the substrate, so its thickness must be larger (> 20 nm) than the dopant diffusion length given our total thermal budget;
- the Si_{1-x}Ge_x channel confines holes, so the Ge fraction must be chosen to avoid the Si cap inversion by the applied gate voltage. If the Ge fraction is low and the gate voltage is high, then holes leave the SiGe quantum well and create a parasitic channel in the Si cap layer, lowering the effective hole mobility. The SiGe channel thickness has to be kept below the critical thickness (< 40 nm) in order to avoid strain relaxation and defect generation.⁶
- the Si cap layer is useful to minimize the scattering on the Si/SiO₂ interface and allows the process to be compatible with the oxidation performed in a standard furnace. However it must be thin enough (> 4 nm) to minimize the average hole mobility reduction by the Si cap inversion when the gate voltage increases.

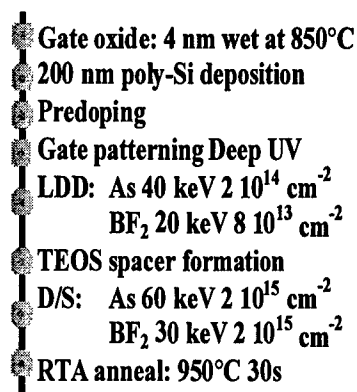


Figure 2. Process sequence.

3. Process sequence

We have fabricated transistors using a standard process sequence, illustrated in Fig. 2. This sequence includes a number of high-temperature steps including: an oxide grown at high temperature (850 °C); two rapid thermal anneals at 950 °C; and other oxidations (gate reoxidation, spacer reoxidation) and as well as spacer deposition carried out at 750 °C. Those anneals could certainly have an influence on the layer quality because they could induce a strain relaxation of the SiGe channel. In addition to the anneals, we used ion implantation at medium dose/low energy and high dose/high energy for the lightly-doped drain (LDD) and source/drain regions respectively. The Ge fraction was 0 (standard Si epitaxy), 15 or 30 percent and the respective layer thicknesses are given in Table 1.

4. Electrical characterization

• Study of defects

Prior to demonstrating the device performance improvement, we investigated the layer quality in terms of defects related to the strain and to the Si/SiGe heterointerface. Two type of techniques were investigated: capacitance and conductance measurements as a function of temperature and frequency; and deep level transient spectroscopy (DLTS) based on an analysis of the drain current transient in the linear regime. Whereas the DLTS measurement does not exhibit any signal with respect to active defects in the SiGe channel, capacitance and conductance measurements show new effects linked to the presence of the SiGe channel (Fig. 3).

Transistors with 15% Ge exhibit a new kink effect near the flat-band voltage and also a new conductance peak. The kink is not due to a freeze out of deep energy levels because this peak is not more pronounced at low T . Instead, we

Split	Ge fraction	Layers thickness (nm)		
		Cap	Channel	Buffer
60 (standard)	0%	—	—	60
6/20 ₁₅ /35	15%	6	20	35
6/20 ₃₀ /35	30%	6	20	35

35	0%	—	—	35
4/20/15	30%	4	20	15
6/10/15	30%	6	10	15
6/15/15	30%	6	15	15
6/20/15	30%	6	20	15
6/20/35	30%	6	20	35
8/20/15	30%	8	20	15

Table 1. Wafer splits used in transistor fabrication.

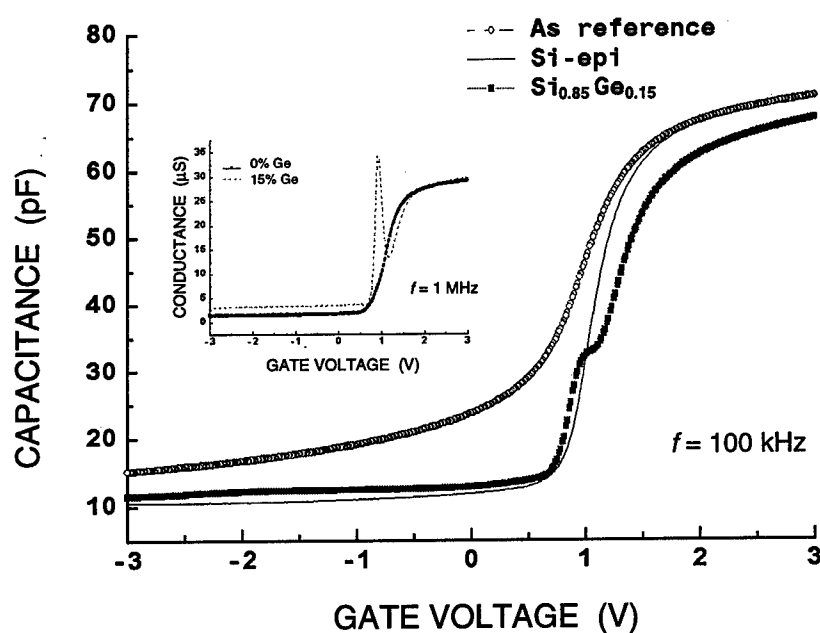


Figure 3. Capacitance and conductance measurement on PMOS transistors with and without SiGe channel.

attribute the feature to the free holes leaving the quantum well. This effect can be used to extract the valence band offset. Actually, if we take the difference between the measured capacitance curve and the theoretical curve corresponding to a case without Ge, then integrating this difference over the voltage range gives the charge.⁷ This charge is directly linked to the valence band offset. The new conductance peak was also used by S. Takagi *et al.*⁸ to extract the valence band offset introduced by the SiGe channel. But this new peak observed in the conductance could also be explained by interface states at the SiO₂ interface. We will see further that this hypothesis is confirmed by physical study and electrical results.

Finally, these measurements made near the flat-band voltage or in the inversion regime allow this technique to explore the heterointerface quality. Actually, the conductance peak at the flat-band voltage and below is clearly linked to the presence of holes at the Si/SiGe interface.

- *Long-channel PMOS performance*

In spite of the high thermal budget, which could relax the strain in the SiGe channel, we present results obtained on PMOS long-channel transistors that show very good performance. Actually, using a Si_{0.7}Ge_{0.3} channel we obtain a gain in hole mobility (at $T = 300$ K), extracted by the "split $C(V)$ method", as large as 110% with respect to the standard architecture (Fig. 4). This gain manifests itself in increased drain current of the long-channel PMOS transistors: 185% at room temperature in the best case and 370% at 85 K (Fig. 4).

According to the DLTS characterization, these very good results are due to hole conduction within a layer without many defects. The mobility and the drain current gain are comparable with results reported elsewhere.²⁻⁵ This mobility improvement is due to the hole effective mass lowering, to the strong reduction in

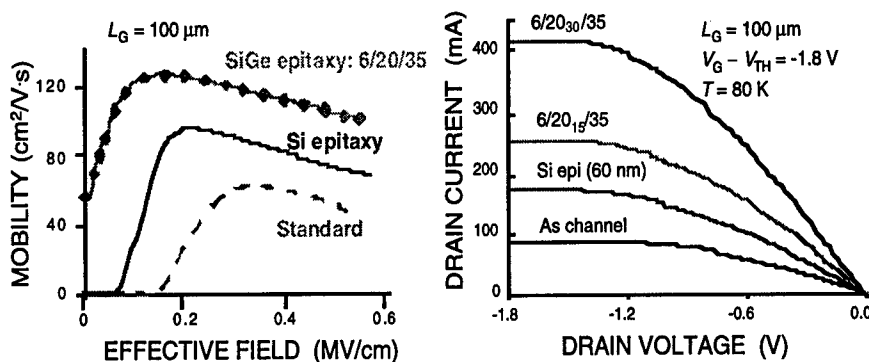


Figure 4. Hole mobility improvement in the strained Si_{0.7}Ge_{0.3} channel vs. Si epitaxy and standard transistor cases (left); drain current gain of long-channel PMOS transistors at $T = 80$ K (right).

ionized impurity and interface roughness scattering. However, it is important to note that the integration of a SiGe channel gives, in many cases, a double conductance peak. One peak corresponds to the SiGe channel inversion and the second to a parasitic inversion channel in the Si cap.⁷ This effect must be avoided if this structure is to be compatible with dynamic device specifications (stability required). Fortunately, we can optimize the layer thickness and shift the Si cap inversion effect towards a value of V_G that exceeds the supply voltage. In this case, the second conductance peak will not be a problem.

- *Extraction of valence band offset and scattering parameters*

The measurements reported above, as well as the capacitance measurement performed as a function of the temperature allow the extraction of the valence band offset and some important scattering parameters. The technique used is very simple and straightforward. It can be performed directly on transistors and does not require a special test structure (Hall structure). The only limit is the resolution of the capacitance meter. We focus on the two main factors, ΔE_V and μ_h , determining the capability of the structure to improve the transistor performance.

As can be seen in Fig. 5, the variation of the threshold voltage versus temperature is linear and the regression coefficient is the same for device architectures with 0, 15 or 30 percent Ge in the channel. Using the model developed by Iniewski *et al.*⁹ it is easy to extract the valence band offset. Indeed, by subtracting the threshold voltage for the SiGe channel from the Si threshold voltage and by eliminating the terms depending on the temperature (regression coefficient is identical) we obtain the following formula:

$$\Delta V_{TH}(\text{Si}_x\text{Ge}_{1-x} - \text{Si}_{epi}) = \Delta V_{FB} + \frac{\Delta E_V(\text{Si}_x\text{Ge}_{1-x})}{q} - qN_B\Delta d_{\max}\left(\frac{1}{C_{\text{cap}}}\right).$$

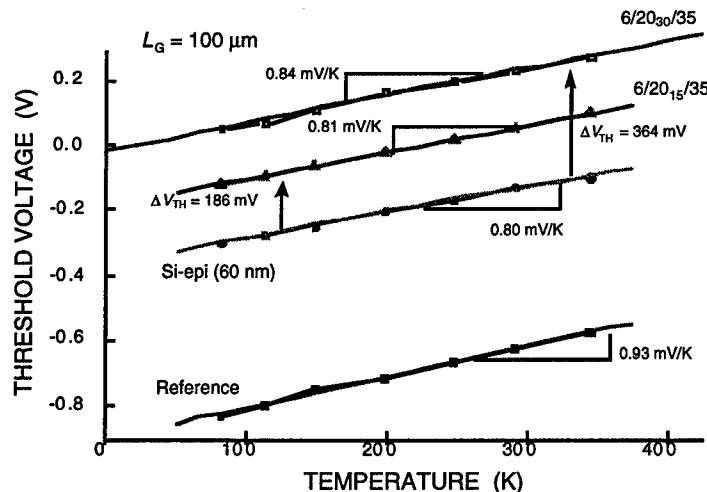


Figure 5. Valence band offset extraction using threshold voltage variation vs. T .

Measuring the flat band voltage (V_{FB}) and knowing the substrate doping level (N_B) as well as the Si cap capacitance (C_{cap}) we finally obtain the valence band offset (ΔE_V).¹⁰ The values we obtain, $\Delta E_V = 160$ meV and 287 meV for $Si_{0.85}Ge_{0.15}$ and $Si_{0.7}Ge_{0.3}$ respectively, are in good agreement with the values obtained on the same SiGe layers using the Hall technique.¹¹

Concerning scattering parameters the following technique is used. The first step is the mobility extraction by the split $C(V)$ method as a function of the temperature. The second step consists of taking the difference between the inverse of the Si channel mobility and the inverse of the SiGe channel one. According to Matthiessen's rule, the difference ($\mu_{SiGe}^{-1} - \mu_{Si}^{-1}$) is proportional to a sum of inverse mobilities linked to the SiGe channel and the Si/SiGe interface. Figure 6 presents a way to extract SiGe phonon parameters using the model developed in Ref. 12, where the inverse of the phonon mobility is defined as $\mu_{ph}^{-1} = \gamma \epsilon_{eff}^{1/3} + \delta$. In Fig. 6, we can clearly see that the variation of the difference ($\mu_{SiGe}^{-1} - \mu_{Si}^{-1}$) in the range of high T is linearly dependent on the effective field to the 1/3 power. This variation reveals a phonon-like behavior that we attribute to phonons in SiGe. Such a technique was also used for low temperature where we have determined scattering parameters due to the Si/SiGe roughness.

5. Short-channel devices

Previous results obtained on long-channel PMOS transistors have given very high gain using SiGe strained channels. However, this gain also has to be available for very short transistors (sub-0.18 μm) for both NMOS and PMOS transistors. In addition, the short-gate transistors must meet the specifications in terms of leakage current, threshold voltage, etc.

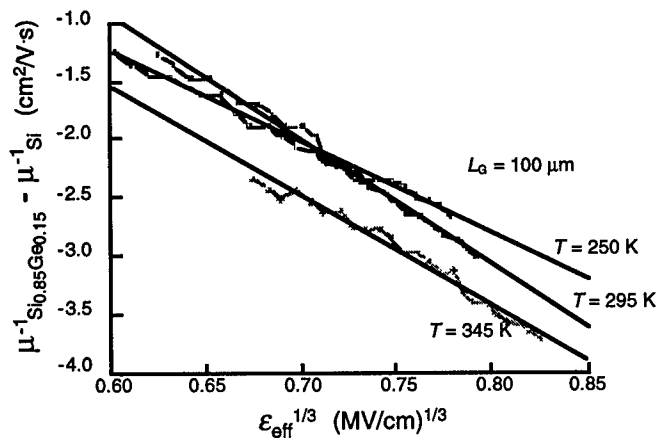


Figure 6. Example of scattering parameter extraction.

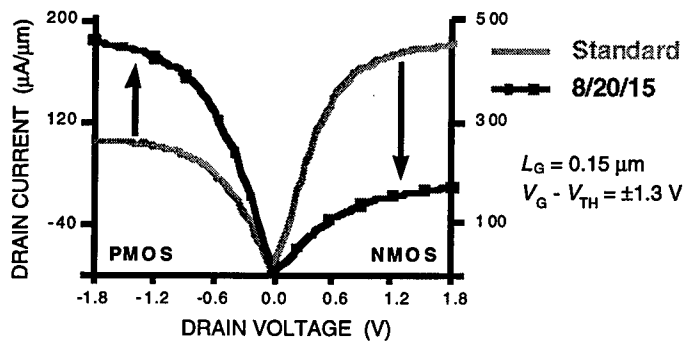


Figure 7. NMOS and PMOS short channel operation ($L_G = 0.15 \mu\text{m}$).

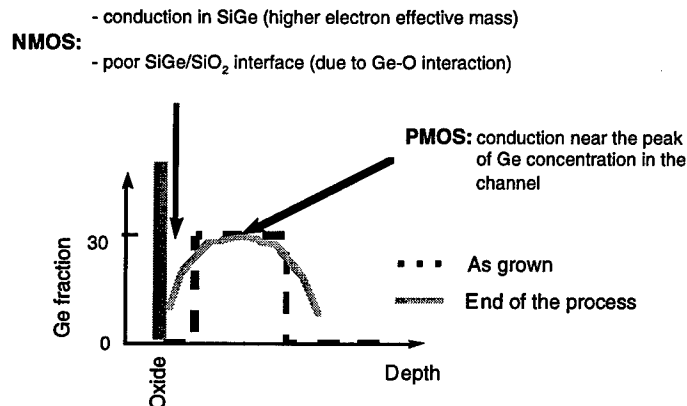


Figure 8. A possible explanation for the asymmetry between NMOS and PMOS.

We have fabricated devices using the 0.15- μm full CMOS process. Figure 7 shows again the performance enhancement obtained in the PMOS device (75%), but also shows the strong degradation of the NMOS transistors. This degradation, as well as the asymmetry observed between NMOS and PMOS devices, can be explained¹³ by the broadened Ge profile at the end of the process. First, the electron longitudinal effective mass is higher in SiGe than in Si,¹⁴ leading to lower electron mobility and hence a lower drain current. Second, in the NMOS device the conduction occurs close to the Si/SiO₂ interface, which becomes an SiGe/SiO₂ interface after the Ge outdiffusion from the SiGe channel. In contrast, in the PMOS device the conduction occurs at the maximum of the Ge concentration sufficiently far away from the poor SiGe/SiO₂ interface (see Fig. 8).

Another critical aspect is the leakage current observed in short channels.¹³ We believe this leakage is due to the damage introduced during the LDD implantation.



Figure 9. SEM cross-sectional photograph of the NMOS transistor.

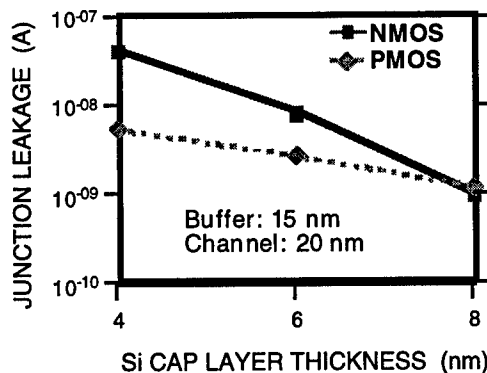


Figure 10. Junction leakage current vs. Si cap layer thickness.

Scanning electron microscopy performed after chemical delineation reveals regions full of defects at the channel edges — see Fig. 9. This region extends 70 nm under the gate electrode in the PMOS device. In the NMOS device, as shown below, the extension of this region is larger due to the arsenic implantation. In addition, electrical results have confirmed¹³ the origin of such leakage current. This region full of defects was not found with DLTS and $C(V)$ measurements because these two techniques were performed on very long capacitors. The presence of defects also explains why we have a big reduction of the average hole mobility (185% gain in long-channel drain current and only 40% in short-channel drain current) in the PMOS device. For the NMOS device, this region full of defects reduces the average mobility and this effect is added to the reduction of the mobility due to the conduction in the SiGe. Fortunately, this leakage level decreases when the Si cap layer thickness increases, as shown in Fig. 10. When the Si cap thickness increases from 4 nm up to 8 nm the junction leakage current decreases by one decade. In this case the leakage is acceptable (1 nA) for the next transistor generation. This figure also shows that the NMOS device is more sensitive to the Si cap thickness because of the more extended defect zone in this case.

6. Conclusions

We have first described the structure used in order to improve the hole mobility in PMOS transistors with the use of strained $\text{Si}_{1-x}\text{Ge}_x$ channels. The three parameters to adjust are the thickness of the Si buffer, the SiGe channel, and the Si cap layers. However these thicknesses have to take into account the NMOS and PMOS short-channel operation and the leakage current for a possible integration within a full CMOS standard process. Nevertheless, this paper shows the big potential for the use of a SiGe channel in terms of mobility and drain current gain obtained in long-channel PMOS devices. In addition, measurements of the threshold voltage and the mobility extracted from the split $C(V)$ method as a function of the temperature comprise a new method to extract the valence band offset and the scattering parameters related to the SiGe phonons and to the Si/SiGe heterointerface. We have also presented results on short channel NMOS and PMOS transistors. We always measure a good gain in drain current in the PMOS case and we have clearly identified the origin of the degradation mechanism for NMOS operation and of the junction leakage current. Fortunately, this degradation (NMOS and leakage) can be avoided by choosing an improved architecture (i.e. leakage reduction for thicker Si cap). This architecture has to realize the best trade-off between the PMOS hole mobility improvement (thin Si cap to reduce parasitic channel), the NMOS operation (thick Si cap to avoid electron conduction in SiGe), the junction leakage (thick Si cap), and the short channel effects (thin Si buffer to control the channel by the ground plane effect).

Finally, the structures studied are a well-adapted approach, and if we slightly modify the process and adapt the architecture, it will be possible to boost the performance for the future generation of transistors.

References

1. R. People, "Physics and applications of $\text{Ge}_x\text{Si}_{1-x}$ strained layer heterostructures," *IEEE J. Quantum Electron.* **22**, 1696 (1986).
2. P. M. Garone, V. Venkataraman, and J. C. Sturm, "Hole confinement in MOS gated $\text{Ge}_x\text{Si}_{1-x}/\text{Si}$ heterostructures," *IEEE Electron Dev. Lett.* **12**, 230 (1991).
3. S. Verdonckt-Vandebroek, E. F. Crabbé, B. S. Meyerson, *et al.*, "SiGe channel heterojunction p -MOSFETs," *IEEE Trans. Electron Dev.* **41**, 90 (1994).
4. S. P. Voinigescu, C. A. T. Salama, J. P. Noël, and T. I. Kamins, "Optimized Ge channel profiles for VLSI compatible Si/SiGe p -MOSFETs," *Tech. Digest IEDM* (1994), p. 369.
5. B. R. Cyca, K. G. Robins, and N. G. Tarr, "Hole confinement and mobility in heterostructure Si/Ge/Si p -channel MOSFETs," *J. Appl. Phys.* **81**, 8079 (1997).
6. R. People and J. C. Bean, "Calculation of critical layer thickness versus lattice mismatch for $\text{Ge}_x\text{Si}_{1-x}$ strained layer heterostructures," *Appl. Phys. Lett.* **47**, 322 (1985).

7. J. Alieu, A. Souifi, G. Brémond, P. Bouillon, and T. Skotnicki, "Electrical characterisation of Si_{1-x}Ge PMOS channel by admittance spectroscopy," *J. Vac. Sci. Technol. B* **16**, 1675 (1998).
8. S. Takagi, J. L. Hoyt, K. Rim, J. J. Welser, and J. F. Gibbons, "Evaluation of the valence band discontinuity of $\text{Si}/\text{Si}_{1-x}\text{Ge}_x/\text{Si}$ heterostructures by application of admittance spectroscopy to MOS capacitors," *IEEE Trans. Electron Dev.* **45**, 494 (1998).
9. K. Iniewski, S. Voinigescu, J. Atcha, and C. A. T. Salama, "Analytical modelling of threshold voltages in p -channel $\text{Si}/\text{SiGe}/\text{Si}$ MOS structures," *Solid State Electronics* **36**, 775 (1993).
10. J. Alieu, R. Gwoziecki, A. Souifi, *et al.*, "Examining the potential of SiGe epitaxial channels for CMOS," *Proc. ISDRS* (1997), p. 501.
11. L. Garchery, Ph. D. Thesis, France Telecom CNET Grenoble (1996)
12. C.-L. Huang and G. S. Gildenblat, "Measurements and modelling of the n -channel MOSFET inversion layer mobility and device characteristics in the temperature range 60–300K," *IEEE Trans. Electron Dev.* **37**, 1289 (1990).
13. J. Alieu, P. Bouillon, R. Gwoziecki, *et al.*, "Optimisation of $\text{Si}_{0.7}\text{Ge}_{0.3}$ channel heterostructures for 0.15/0.18 μm CMOS process," *Proc. ESSDERC* (1998), p. 144.
14. S. M. Sze, *Physics of Semiconductors Devices*, 2nd ed., New York: Wiley, 1981.

Device Implications of Strain Relaxation in Semiconductor Microstructures

C. D. Akyüz

Department of Physics, Brown University, Providence, RI 02912

H. T. Johnson, A. Zaslavsky, and L. B. Freund

Division of Engineering, Brown University, Providence, RI 02912

D. A. Syphers

Department of Physics, Bowdoin College, Brunswick, ME 04011

1. Introduction

The quest for new and better (faster, smaller, more efficient) semiconductor electronic devices to meet the demands of current and future technology has been pushing the fabrication limits for many decades.¹ Many advanced devices involve epitaxial heterojunctions that make it possible to tailor precisely the areas of carrier localization, built-in electric fields, *etc.* The success of modern epitaxy and bandgap engineering has encouraged researchers to develop epitaxial techniques for lattice-mismatched materials. Strained layer epitaxy has increased the number of available heterojunctions, as in Si/SiGe heterojunction bipolar transistors (HBTs) or InP-based strained quantum well lasers. The availability of strained Si/SiGe heterostructures has been particularly significant as it has introduced band structure engineering to the dominant silicon technology.

The properties of homogeneous biaxial strain arising from lattice mismatch and its effects on the electronic properties of semiconductor thin films are well understood.²⁻⁴ Biaxial compression can be broken up into a hydrostatic compression and a uniaxial expansion as illustrated in Fig. 1. The hydrostatic term changes the bandgap but shifts the valence bands uniformly, leaving unchanged the heavy-hole (HH) and light-hole (LH) band-edge degeneracy of technological semiconductors. The uniaxial expansion introduces an interaction between the bands, lifting the HH-LH band-edge degeneracy as shown in Fig. 1.

A number of novel microelectronic devices based on strained heterostructures have been proposed or reported in the literature. In Si/SiGe field-effect transistors, strain-controlled band alignment confines the carriers in the channel.⁵ Si/SiGe HBT designs (shown in Fig. 2) use the heterojunction to provide a built-in electric field and to tune the effective mass of carriers in the narrow, strained base.⁶ In multiple⁷⁻⁹ and single quantum well¹⁰ lasers, the strain-induced HH-LH splitting is optimized for device efficiency; the control over the strain-induced HH-LH splitting also helps to build polarization-insensitive laser amplifiers.¹¹ Other strained devices include phototransistors¹² and infrared detectors.¹³

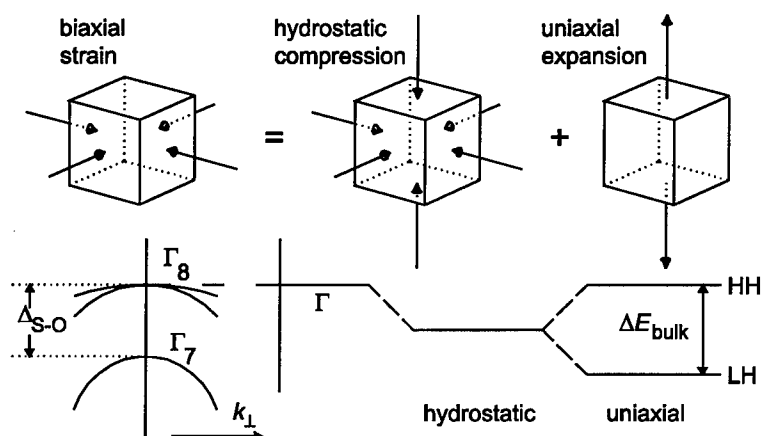


Figure 1. The effect of biaxial strain on the valence band structure. The uniaxial component lifts the valence band degeneracy ($\Delta E_{\text{bulk}} \approx 42 \text{ meV}$ in $\text{Si}_{0.75}\text{Ge}_{0.25}$ on Si).

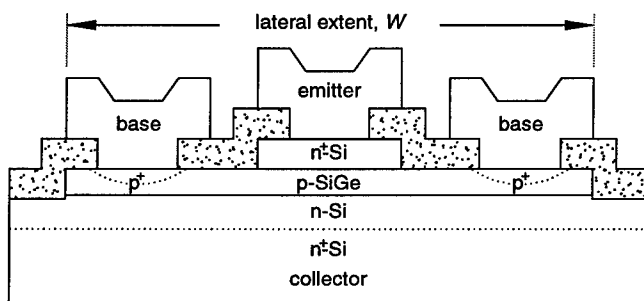


Figure 2. Schematic cross-section of an *nnp* Si/SiGe/Si heterojunction bipolar transistor. In high-speed devices, the lateral extent of the strained base may need to become submicron.

In current semiconductor device technology, device efficiency is increased by reducing the lateral device size (see Fig. 2). Reduction in size leads to lower power consumption and faster switching, so critical dimensions of modern semiconductor devices are downscaling rapidly into the deep submicron range. When semiconductor structures are etched in strained material, strained layers relax by lateral displacement of the free surface resulting in a nonuniform strain distribution. Previously, the effects of nonuniform strain were investigated in III-V microstructures fabricated by stressor patterning¹⁴ or epitaxial regrowth.^{15,16} More recently, tunneling spectroscopy of technologically interesting Si/SiGe microstructures has revealed significant changes in their electronic characteristics due to size-induced strain relaxation.¹⁷⁻¹⁹

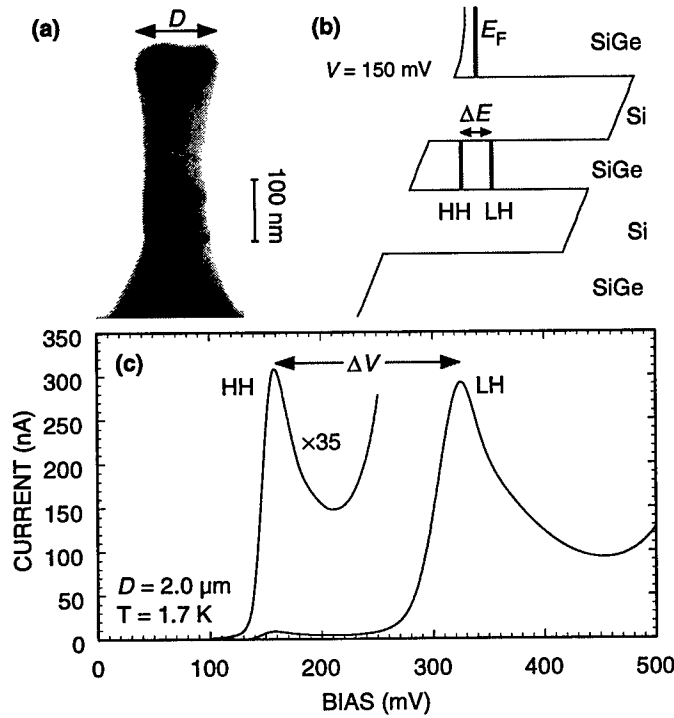


Figure 3. (a) SEM micrograph of a representative double-barrier nanostructure with lateral dimension $D = 0.15$ μm . (b) Self-consistent potential profile of the Si/Si_{0.75}Ge_{0.25} double-barrier active region at $V = 150$ mV. (c) Resonant tunneling $I(V)$ of a large $D = 2$ μm device at 1.7 K, with the HH peak magnified x35 for clarity.

2. Strain relaxation in Si/SiGe microstructures

Lately, numerical methods have been applied to estimate the effects of size-induced strain relaxation in semiconductor structures of submicron sizes with built-in strain.^{18,20,21} We have employed finite-element techniques to calculate the strain distribution, and resonant tunneling (RT) measurements to probe the effects of size-induced strain relaxation on the electronic properties in Si/Si_{0.75}Ge_{0.25} double barrier RT diodes with lateral dimensions in the $0.1 \leq D \leq 2$ μm range. The details of the structure are published elsewhere.¹⁷ Here it suffices to note that the active region consists of an undoped 35 Å Si_{0.75}Ge_{0.25} quantum well between two 45 Å Si barriers, which in turn are surrounded by p -Si_{0.75}Ge_{0.25} electrodes. The processing has involved conventional e-beam lithography and metallization techniques for the Ti/Al top contacts, followed by reactive ion etching of the mesa (see Fig. 3(a) for an SEM photograph of a $D = 0.15$ μm device), SiO₂-based planarization, etchback, and contact pad deposition. When a bias V is applied between the top contact and the substrate, a tunneling current $I(V)$ flows through the quantized 2D hole subbands in the well subject to the usual energy E and

transverse momentum k_{\perp} conservation rules.²² Each 2D subband gives rise to a resonant peak in the $I(V)$. For our double-barrier structure, the well contains two subbands that can be classified as HH and LH according to the dominant character of the hole states at small k_{\perp} , so the $I(V)$ contains two peaks. Self consistent calculations are used to convert the peak separations ΔV , in the tunneling current to strain-dependent HH-LH subband separation ΔE in the well.¹⁷ The calculated potential distribution and the characteristic $I(V)$ of a large $D = 2 \mu\text{m}$ device at $T = 1.7 \text{ K}$ is shown in Figs. 3(b) and 3(c), respectively. The lower bias $I(V)$ peak at $V = 150 \text{ mV}$ shown in Fig. 3(c) corresponds to tunneling through the HH subband, while peak at $V = 315 \text{ mV}$ corresponds to tunneling through the LH subband — the corresponding HH-LH subband separation ΔE agrees exactly with the subband spectrum calculated using a 6X6 Luttinger-Kohn Hamiltonian.¹⁷

When a submicron mesa is etched out of a strained Si/SiGe double-barrier RT material, the strained layers relax by sidewall expansion. We have calculated the strain components of cylindrical Si/Si_{0.75}Ge_{0.25} pillars of various D by finite-element simulations based on a linear elastic model,²³ where cylindrical pillars with fully strained Si_{0.75}Ge_{0.25} layers were allowed to relax to a state of minimum energy. The geometry of the calculation is illustrated in the inset of Fig. 4, where the lateral sidewall displacement is magnified for clarity. The dominant effect is

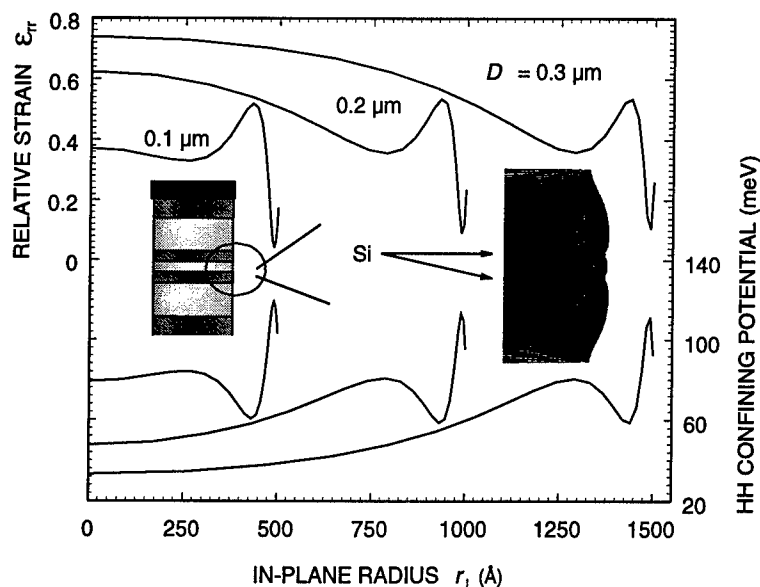


Figure 4. The top curves show the calculated radial strain component ϵ_r as a function of radius r_{\perp} for $D = 0.1, 0.2$, and $0.3 \mu\text{m}$ devices on the mid-plane of the Si_{0.75}Ge_{0.25} well (full biaxial strain corresponds to $\epsilon_r = 1$). Inset shows the magnified displacement of the finite-element mesh corresponding to a section of the active region near the sidewall. The bottom curves are the corresponding in-plane confining potentials for the HH states as function of r_{\perp} . Note that in the $D = 0.1 \mu\text{m}$ device, the ground state is confined to a narrow ring-like region at the perimeter.

the relaxation of the in-plane strain component ϵ_{rr} , which is plotted as a function of r_{\perp} on the mid-plane of the $\text{Si}_{0.75}\text{Ge}_{0.25}$ well for $D = 0.1\text{--}0.3\text{ }\mu\text{m}$ (top curves in Fig. 4). In the calculations, strain is normalized to full biaxial strain. The radial component of the strain decreases gradually with r_{\perp} from the center of the pillar ($r_{\perp} = 0$) with a strong inhomogeneous region of increasing strain near the surface ($r_{\perp} = D/2$). A ring-like region with $\epsilon_{rr} \approx 0.5$ and a radial extent of $\sim 100\text{ }\text{\AA}$ runs around the perimeter of the structure for all D .

In larger devices, $D \geq 0.3\text{ }\mu\text{m}$, the strain distribution is similar to Fig. 4, with the strongly inhomogeneous region becoming a small perturbation to the approximately uniformly relaxed inner core. The amount of relaxation in the inner core depends on D and the lattice mismatch ($\sim 1\%$ in $\text{Si}/\text{Si}_{0.75}\text{Ge}_{0.25}$): at the center of the pillar we find the radial strain component ϵ_{rr} relaxes $\sim 15\%$ when $D = 0.75\text{ }\mu\text{m}$ and $\sim 30\%$ when $D = 0.3\text{ }\mu\text{m}$. Even in these larger devices, the inhomogeneously strained regions near the sidewalls can affect the performance of devices in which current crowding near the edges is significant — e.g. stripe geometry HBTs as in Fig. 2. In smaller nanostructures ($D < 0.3\text{ }\mu\text{m}$ for $\text{Si}/\text{Si}_{0.75}\text{Ge}_{0.25}$), the inhomogeneity of the strain relaxation becomes significant throughout the structure, as seen in Fig. 3.

The average strain relaxation in the central core of submicron RT pillars can be extracted from the resonant peak positions in $I(V)$ measurements.¹⁷ Figure 5 presents $I(V)$ characteristics for $D = 1\text{--}0.25\text{ }\mu\text{m}$ devices measured at $T = 1.7\text{ K}$. The data exhibit a consistent shift in the HH and LH peaks towards each other

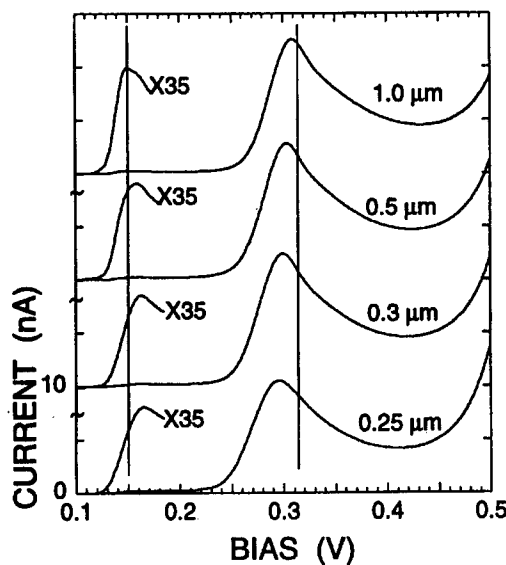


Figure 5. $I(V)$ characteristics with nominal lateral diameters of $D = 1.0, 0.5, 0.3$, and $0.25\text{ }\mu\text{m}$ at 1.7 K . The current scale corresponds to the smallest device. Other curves are rescaled and shifted, and the HH resonant peaks are magnified (X35) for clarity. The vertical lines indicate the HH and LH peak positions in large devices.

with decreasing D . The overall lineshape of the resonant current is retained with very good peak-to-valley ratios indicating minimal surface damage. The size dependent changes in the $I(V)$ can be unambiguously attributed to strain relaxation, which reduces the HH-LH subband separation ΔE of Fig. 3(b).

In bulk $\text{Si}_{1-x}\text{Ge}_x$ strained to Si, the HH-LH band-edge separation is given by:⁴

$$\Delta E_{\text{bulk}}(x) = \frac{[3\xi(x) + \Delta(x)] - \sqrt{9\xi^2(x) + \Delta^2(x) - 2\xi(x)\Delta(x)}}{2}, \quad (1)$$

where $\Delta(x)$ is the interpolated spin-orbit splitting and $\xi(x)$ is the strain energy due to lattice mismatch. The values $\Delta(0.25) = 106$ meV and $\xi(0.25) = 31$ meV give $\Delta E_{\text{bulk}}(0.25) = 42$ meV for fully strained material, while any relaxation of the biaxial strain reduces $\xi(x)$ and hence ΔE_{bulk} . In a quantum well the situation is more complex and the change in ΔE arising from strain relaxation must be evaluated numerically.^{18,19}

In Fig. 6, we plot the calculated strain at the center and average in-plane strain in the quantum well versus lateral device size for two geometries. In Fig. 6(a), the device is a cylindrical pillar with lateral diameter D , directly comparable to our experimental data. In Fig. 6(b), the device is a rectangular mesa of lateral width W much smaller than the length — a geometry relevant to HBTs and stripe-geometry lasers for which similar RT measurements have been reported in Ref. 19. As in Fig. 4, the strain components ϵ_{rr} and ϵ_{xx} are normalized to full biaxial strain. The experimental data points in Fig. 6(a) are the average strain values extracted from the data in Fig. 5 using a self-consistent calculation to convert ΔV into ΔE and then a 6X6 Luttinger-Kohn Hamiltonian to convert ΔE in the quantum well to strain relaxation. Our data are in good agreement with finite-element calculations and prove the strain relaxation in submicron structures to be a large effect.

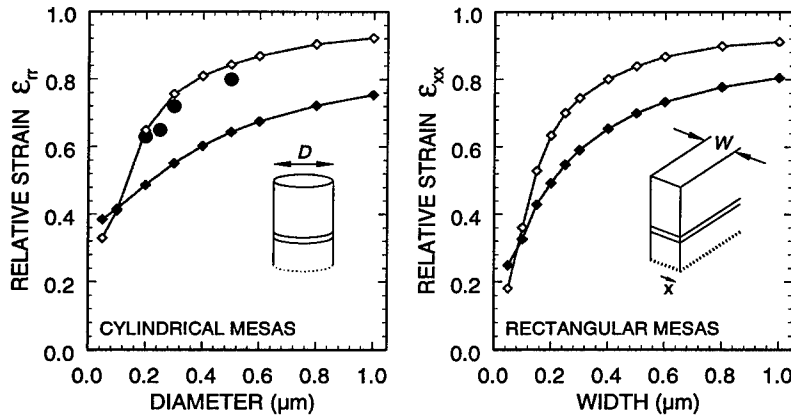


Figure 6. (a) Calculated radial strain ϵ_{rr} at the center ($r_L = 0$, open diamonds) and average radial strain (filled diamonds) in the well of a cylindrical RT structure vs. D . (b) Calculated in-plane strain component ϵ_{xx} of a stripe geometry structure vs. width W . The filled circles in (a) are data extracted from our tunneling measurements.

3. Inhomogeneous strain relaxation

The disparity between the strain at the center and the average strain in Fig. 6 is a measure of the inhomogeneity in the strain. The inhomogeneous regions near the sidewalls predicted for devices with very small diameters $D < 0.25 \mu\text{m}$ (see Fig. 3) manifest themselves in transport measurements. Resonant tunneling $I(V)$ data presented in Figs. 7(a,b) for $D = 0.25$ and $0.2 \mu\text{m}$ devices show a consistent development of quasi-periodic fine structure in the HH peak. The separation between features is ~ 7 mV, corresponding to an energy separation of ~ 2 meV. For these values of D , the LH peak lineshape does not change from the large device $I(V)$ of Fig. 3, apart from the shift in peak position due to average strain relaxation. The lack of observable fine structure in the LH peaks is consistent with the heavier in-plane mass predicted for light holes in a quantum well. Figure 7(c) shows both the HH and LH $I(V)$ peaks of a $D = 0.1 \mu\text{m}$ device. Here the HH peak has changed qualitatively into steps on a rising current background. The step separation is ~ 35 mV, corresponding to ~ 8 meV energy separation, and the higher step is split into two closely spaced features. The LH resonant peak at $D = 0.1 \mu\text{m}$ has developed a fine structure quite similar to that exhibited by HH peaks in larger devices. It is important to note that for the device sizes we have used, the lateral quantization energy scale $\hbar^2/2m^*D^2 \ll 1$ meV. The fine structure is reproducible upon temperature cycling, making tunneling through impurity states an unlikely explanation.

We attribute the resonant peak fine structure to lateral quantization in inhomogeneous strain-induced potentials.¹⁸ We converted the inhomogeneous strain distribution (top curves in Fig. 4) to a corresponding lateral confining potential for HH states by retaining only the dominant ϵ_r term that is most affected by strain relaxation. We then calculate the HH subband energy at $k_{\perp} = 0$ as a function of local strain $\epsilon_r(r_{\perp})$ by the usual 6X6 Luttinger-Kohn Hamiltonian.³ The resulting lateral potentials for HH states are shown on the bottom of Fig. 4. The regions with larger ϵ_r are local potential minima for HH states. Accordingly, the lateral confining potential for $D > 0.2 \mu\text{m}$ structures looks approximately harmonic, with a ring-like perturbation at the perimeter, while for the smaller $D = 0.1 \mu\text{m}$ structure it is the ring that confines the ground state. The effective height of the lateral confinement is ~ 15 meV. Analogous strain gradients appear in the SiGe emitter region (see inset of Fig. 4), but there, the lateral potential is screened by the large density of holes. Consequently, the observed fine structure arises principally from the lateral quantization in the well, where the density of dynamically stored holes is negligible.^{18,24} Given the confining HH potentials of Fig. 4, one can estimate the lateral confinement energies using interpolated HH effective masses at $k_{\perp} = 0$. The result is that for the $D = 0.2 \mu\text{m}$ structure, the lowest-lying states are confined near $r_{\perp} = 0$ by the approximately parabolic potential and separated by ~ 2 meV. Conversely, in the $D = 0.1 \mu\text{m}$ device, the two lowest-lying states are confined to a narrow ring at the perimeter and their energy separation is on the order of ~ 10 meV. These estimates of lateral energy quantization are in excellent agreement with the observed HH peak lineshapes, which evolve from relatively weak quasiperiodic fine structure at $D = 0.25 \mu\text{m}$ to

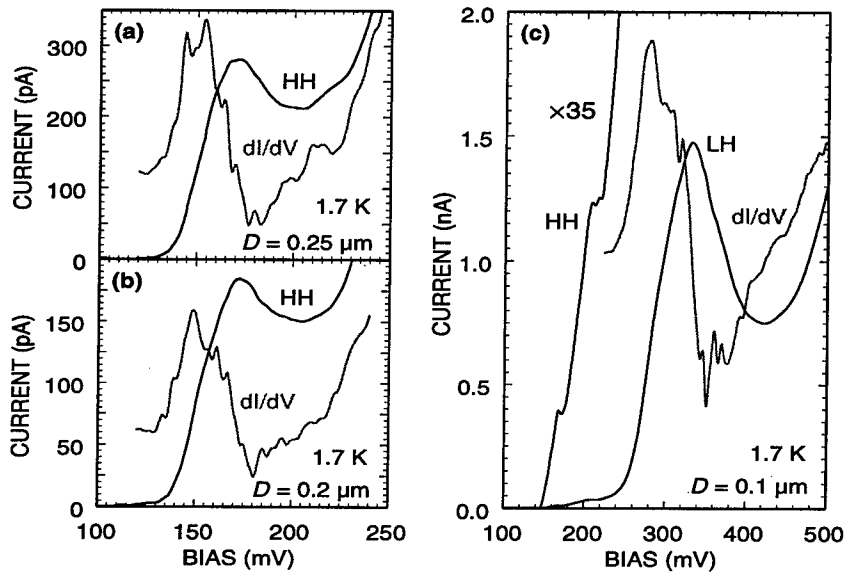


Figure 7. $I(V)$ and dI/dV characteristics of $D \leq 0.25 \mu\text{m}$ devices at $T = 1.7$ K. The HH $I(V)$ peaks of $D = 0.25 \mu\text{m}$ (a) and $D = 0.2 \mu\text{m}$ (b) devices exhibit reproducible quasiperiodic fine structures, while the LH peaks remain smooth. When $D = 0.1 \mu\text{m}$ (c), the HH peak changes qualitatively into two well-resolved steps, while the LH peak begins to exhibit a fine structure.

two well-resolved current steps at $D = 0.1 \mu\text{m}$ (consistent with the sharp outer ring potential).¹⁸ The inhomogeneous strain relaxation may be used to fabricate scientifically interesting quantum structures like quantum rings that would have fascinating magnetic properties. At the same time, devices that rely on uniform biaxial strain may be adversely affected by inhomogeneous strain relaxation once they are scaled down into the deep submicron regime.

4. Conclusions

We have investigated the effects of size-induced strain relaxation on the electronic properties of strained heterostructure devices. Resonant tunneling measurements were used to probe the strain relaxation and associated changes in the electronic properties of individual $p\text{-Si/Si}_{0.75}\text{Ge}_{0.25}$ double-barrier resonant tunneling diodes of diameters in the $0.1 \leq D < 2 \mu\text{m}$ range. In the tunneling current we observed an overall shift in heavy and light-hole peaks towards each other as the device diameter is decreased in the $2 > D > 0.25 \mu\text{m}$ range due to strain relaxation in the devices. Calculations based on finite-element techniques confirm the magnitude of size-induced strain relaxation and the results are in good agreement with the data.

In smaller devices with $D \leq 0.25 \mu\text{m}$, the $I(V)$ exhibit a fine structure due to lateral quantization in an inhomogeneous strain-induced potential. We correlated the measured HH fine structure with the confining potentials corresponding to the nonuniform strain distributions predicted by finite-element analysis. Our measurements prove tunneling to be a useful spectroscopic probe for strain effects in individual nanostructures, confirm the effectiveness of finite-element calculations even at these deep submicron length scales, and point out the surprisingly large influence size-induced strain relaxation will exert on the electronic properties of strained submicron devices.

5. Acknowledgments

We thank A. Schwartzman for microscopy assistance. A. Z. and C. D. A. have been supported by an NSF Career award (DMR-9702725), the ONR Young Investigator Program (N00014-95-1-0729), and a Sloan Foundation fellowship; while H. T. J. and L. B. F. acknowledge support by ONR (N00014-95-1-0239) and the MRSEC Program of the NSF (DMR-9632524).

References

1. S. M. Sze, ed., *Modern Semiconductor Device Physics*, New York: Wiley, 1998.
2. G. L. Bir and G. E. Pikus, *Symmetry and Strain-Induced Effects in Semiconductors*, New York: Wiley, 1974.
3. J. M. Luttinger and W. Kohn, "Motion of electrons and holes in perturbed periodic fields," *Phys. Rev.* **97**, 869 (1955).
4. R. People, "Indirect bandgap of coherently strained $\text{Ge}_x\text{Si}_{1-x}$ bulk alloys on $\langle 001 \rangle$ silicon substrates," *Phys. Rev. B* **32**, 1405 (1985).
5. H. Daembkes, H.-J. Herzog, H. Jorke, H. Kibbel, and E. Kasper, "The n -channel SiGe/Si modulation-doped field-effect transistor," *IEEE Trans. Electron Dev.* **33**, 633 (1986).
6. S. S. Iyer, G. L. Patton, J. M. C. Stork, B. S. Meyerson, and D. L. Harnage, "Heterojunction bipolar transistors Using Si-Ge Alloys," *IEEE Trans. Electron Dev.* **36**, 2043 (1989).
7. D. Nicholls and P. Bhattacharya, "Differential gain in InP-based strained layer multiple quantum well lasers," *Appl. Phys. Lett.* **61**, 2129 (1992).
8. H. Lu, C. Blaauw, T. Makino, and M. Gallant, "High temperature operation of $1.3 \mu\text{m}$ ridge waveguide lasers using lattice matched and strained multiple quantum wells," *Appl. Phys. Lett.* **64**, 2761 (1994).
9. M. Itoh, H. Sugiura, H. Yasaka, Y. Kondo, and K. Kishi, "Differential gain and threshold current of $1.3 \mu\text{m}$ tensile-strained InGaAsP multiquantum well buried-heterostructure lasers grown by metalorganic molecular beam epitaxial growth," *Appl. Phys. Lett.* **72**, 1553 (1998).

10. F. Agahi, A. Baliga, K. M. Lau, and N. G. Anderson, "Tensile strained barrier GaAsP/GaAs single quantum well lasers," *Appl. Phys. Lett.* **68**, 3778 (1996).
11. M. Jorna, H. Horikawa, C. Q. Xu, *et al.*, "Polarization insensitive semiconductor laser amplifiers with tensile strained InGaAsP/InGaAsP multiple quantum well structure," *Appl. Phys. Lett.* **62**, 121 (1993).
12. M. Gisoni, O. Sjölund, and A. Larsson, "Comparison of partially relaxed InGaAs/GaAs based high performance phototransistors," *Appl. Phys. Lett.* **69**, 1773 (1996).
13. F. Y Huang, X. Zhu, M. O. Tanner, and K. L. Wang, "Normal incidence strained-layer superlattice $\text{Ge}_{0.5}\text{Si}_{0.5}/\text{Si}$ photodiodes near $1.3\ \mu\text{m}$," *Appl. Phys. Lett.* **67**, 566 (1995).
14. K. Kash, B. P. Van der Gaag, Derek D. Mahoney, *et al.*, "Observation of quantum confinement by strain gradients," *Phys. Rev. Lett.* **67**, 1326 (1991).
15. D. Gershoni, J. S. Weiner, S. N. Chu *et al.*, "Optical transitions in quantum wires with strain-induced lateral confinement," *Phys. Rev. Lett.* **65**, 1631 (1990).
16. T. Arakawa, S. Tsukamoto, Y. Nagamune, *et al.*, "Fabrication of InGaAs strained quantum wire structures using selective-area metalorganic chemical vapor deposition growth," *Jpn. J. Appl. Phys.* **2** **32**, L1377 (1993).
17. A. Zaslavsky, K. R. Milkove, Y. H. Lee, B. Ferland, and T. O. Sedgwick, "Strain relaxation in silicon-germanium microstructures observed by resonant tunneling spectroscopy," *Appl. Phys. Lett.* **67**, 3921 (1995).
18. C. D. Akyüz, A. Zaslavsky, L. B. Freund, D. A. Syphers, and T. O. Sedgwick, "Inhomogeneous strain in individual quantum dots probed by transport measurements," *Appl. Phys. Lett.* **72**, 1739 (1998).
19. P. W. Lukey, J. Caro, T. Zijlstra, E. van der Drift, and S. Radelaar, "Observation of strain-relaxation-induced size effects in *p*-type Si/SiGe resonant tunneling diodes," *Phys. Rev. B* **57**, 7132 (1998).
20. L. De Caro, L. Tapfer, and A. Giuffrida, "Finite size effects in one-dimensional strained semiconductor heterostructures," *Phys. Rev. B* **54**, 10575 (1996).
21. H. T. Johnson, L. B. Freund, C. D. Akyüz, and A. Zaslavsky, "Finite element analysis of strain effects on electronic and transport properties in quantum dots and wires," *J. Appl. Phys.* **84**, 3714 (1998).
22. S. Luryi, "Frequency limit of double-barrier resonant-tunneling oscillators," *Appl. Phys. Lett.* **47**, 490 (1985).
23. These calculations were performed using the ABAQUS finite-element code for mechanical analysis (Hibbitt, Karlsson & Sorensen, Inc., Pawtucket, RI).
24. V. J. Goldman, D. C. Tsui, and J. E. Cunningham, "Resonant tunneling in magnetic fields: evidence for space-charge buildup," *Phys. Rev. B* **35**, 9387 (1987).

The PNP Heterojunction Bipolar Transistor: What Will Be Its Impact in the 21st Century?

S. Ekbote, S. Datta, M. Cahay, and K. Roenker

Dept. of Electrical Engineering, Univ. of Cincinnati, Cincinnati, Ohio 45221

1. Introduction

The glorious days of Si-based technologies are far from over, but with the fast approaching era where Moore's Law will no longer be applicable, there is an urgent need to push state-of-the-art technologies based on different materials towards full-fledged commercial applications. With the turn of the century, we expect to see technologies based on different materials (such as GaAs, GaN, InP, high- T_c superconductors among others) to offer lucrative niche areas with a wide variety of applications spanning the field of telecommunications, microwave applications, and high-speed digital circuits. One device that shows great promise for these commercial applications is the InP-based *pn*p heterojunction bipolar transistor (HBT). In this article, we review the recent experimental and modeling efforts in the design of prototype InP-based *pn*p HBT devices.

2. PNP HBT technology

The development of *pn*p InP-based HBTs has received relatively little attention.¹⁻⁴ While high performance InP-based *npn* HBTs have been widely reported for InAlAs/InGaAs^{5,6} and InP/InGaAs⁷ material systems over the past several years, *pn*p transistors in the InAlAs/InGaAs^{1,3} and InP/InGaAs⁴ material systems have only recently been demonstrated in the laboratory. Their emergence, however, coupled with their demonstrated high gain and high frequency (GHz) performance and the recent successful integration of *pn*p and *npn* devices on the same substrate^{1,2} now make feasible and attractive the development of an InP-based complementary HBT IC technology. The utility and applications of *pn*p bipolar transistors are well known from silicon ICs. They are used as active loads, which provide high gain with reduced parasitics, and are therefore attractive for wide band microwave amplifiers, and as complementary transistors in push-pull amplifiers for power applications.⁸

In recent years the development of *pn*p GaAs-based transistors and a complementary HBT technology in the AlGaAs/GaAs materials system has attracted some interest and seen some progress.⁹ For example, Enquist *et al.* have demonstrated *pn*p transistors with gains of up to 300 at 1.5×10^4 A/cm² with an f_T of 21 GHz and an f_{max} of 23 GHz.¹⁰ Based on these devices, they have demonstrated a low-power, high-speed, complementary HBT-based integrated

injection logic (f^2L) with 65 ps and 13 mW per gate for a speed-power product of 850 fJ. Hill *et al.* and Liu *et al.* have obtained *pn*p HBTs with current gain of 200, an f_T of 23 GHz and an f_{max} of 40 GHz, and demonstrated a push-pull power amplifier at 10 GHz with an output power of 500 mW, 6 dB gain, and 41.8% power-added efficiency.^{11,12}

For the InP-based transistors, there have been only a few reports of *pn*p HBTs. Stanchina *et al.* have obtained current gain of 25, an f_T of 10 GHz and an f_{max} of 27 GHz at 7×10^3 A/cm² for an initial, conservatively designed, single heterojunction InAlAs/InGaAs *pn*p transistor.² More recently, they obtained current gain of 170, an f_T of 13 GHz and an f_{max} of 20 GHz, and successfully demonstrated coplanar monolithic integration of the *pn*p and *npn* transistors using MBE growth on a mesa-patterned substrate. For the related InP/InGaAs HBTs, thus far, there has been only a single report of an operational *pn*p HBT with a peak current gain of 420, an f_T of 10.5 GHz and an f_{max} of 25 GHz.⁴

These preliminary performance results for prototype InP-based *pn*p HBTs are admittedly inferior, but comparable in magnitude, to those obtained for the more mature AlGaAs/GaAs material system, as well as the highly perfected silicon *pn*p transistors employing submicron (0.5 μ m wide) emitters. But there is optimism that significant performance improvements are possible for InP-based HBTs as their design, epitaxy and fabrication are refined. In support of this supposition we note that Hutchby¹³ has theoretically estimated a *pn*p AlGaAs/GaAs HBT to be capable of an f_T of 31 GHz and an f_{max} of 94 GHz for a 1 μ m wide emitter, parameter values comparable to the *npn* HBT. While the *pn*p device structure needs to be optimized for each material system, projections of *pn*p performance comparable to that of the *npn* has been reported by other groups.^{14,15} Given the superior material properties of InGaAs relative to GaAs — higher electron mobility, larger band discontinuities and lower contact resistances — it is anticipated that InP-based *pn*p HBTs will be capable of high frequency performance comparable or superior to that of *pn*p GaAs-based HBTs. This achievement would open the door for the development of a high-speed, InP-based complementary HBT IC technology.

3. Numerical simulations

Recently, we reported some of the first numerical simulations of *pn*p InP/InGaAs and InAlAs/InGaAs HBTs.¹⁶⁻¹⁹ We have investigated the following design aspects of *pn*p HBTs: (1) the influence of the base and collector doping on the current and power gains, on the unity-gain current cutoff frequency and on the maximum frequency of operation;¹⁶ (2) the importance of the hole quasi-Fermi level splitting at the emitter-base junction;¹⁸ (3) the effects of base grading on high-frequency performance;¹⁸ and (4) the importance of the spin-orbit split-off band on the tunneling properties of holes through abrupt heterojunctions.¹⁹

In Ref. 16, we reported the results of *pn*p HBT modeling based on a commercial simulator assuming a drift-diffusion model for carrier transport. We simulated the *pn*p InAlAs/InGaAs HBT of Fig. 1, fabricated and characterized

p+ InGaAs	1×10^{19}	0.135 μ	Emitter Contact
p+ InAlAs	2×10^{18}	0.05 μ	
p InAlAs	8×10^{17}	0.11 μ	Emitter
p-n InGaAlAs		0.03 μ	Graded Layer
n+ InGaAs	7×10^{18}	0.033 μ	Base
p- InGaAs	1×10^{17}	0.25 μ	Collector
p+ InGaAs	5×10^{18}	0.7 μ	Subcollector
i InGaAs			Buffer
i InP			Substrate

Figure 1. Epitaxial structure for *pnp* InAlAs/InGaAs HBT with quaternary emitter-base graded layer.

by researchers at Hughes Research Laboratories.^{1,2} In the simulations, the graded layer at the emitter-base interface was replaced by a set of two quaternary layers of intermediate composition (40% and 60% of GaAs) with thicknesses of 15 nm each and dopings of $8 \times 10^{17} \text{ cm}^{-3}$ and *p* and *n*-type, respectively.

Seen in Fig. 2 are the calculated dc and small signal current gains of the above described device, as well as the experimental results for a $2 \times 10 \mu\text{m}^2$ emitter device for $V_{CE} = -3.5 \text{ V}$. The dc current gain is in reasonable agreement with the experimental results peaking at 170 near a collector current density of 10^4 A/cm^2 . For the ac current gain at 10^4 A/cm^2 , the simulation results are of the order of 250 and nearly flat over several orders of magnitude in the current density. Both the experimental and calculated results fall off above 10^4 A/cm^2 .

The simulation results overestimate both current gains at low collector current densities, perhaps due to the omission of surface recombination in the base in the simulations. Also shown in Fig. 2, the device's unilateral power gain was calculated as a function of the collector current density and found to peak at 38 dB near 10^3 A/cm^2 .

The frequency dependence of the current and power gain were also examined to determine the cutoff frequency (f_T) and maximum frequency of oscillation (f_{max}). Figure 3 shows that there is a satisfactory agreement between the calculated and measured f_T over a wide range of the collector current density. In comparison with the *npn* HBT, the *pnp* HBT is seen to be limited in its gain and high frequency capability (somewhat lower f_T and f_{max}) due to the lower mobility of holes. However, the high electron mobility in the base is expected to somewhat compensate and reduce the base spreading resistance allowing for narrower base widths and lower base doping while maintaining a high f_{max} .¹³⁻¹⁵

We also have developed an analytical model based on a thermionic emission-diffusion model of graded-base *pnp* HBTs, taking into account the effects of hole

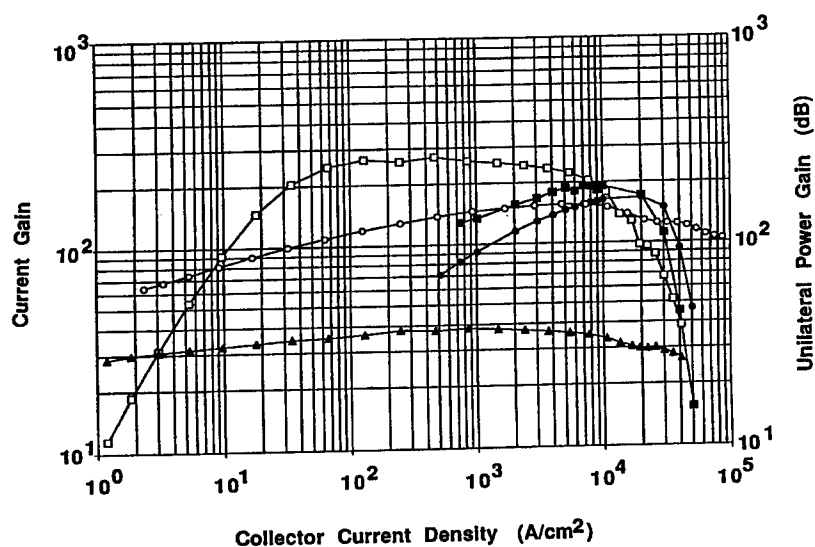


Figure 2. Calculated dc (open circles) and small signal (open squares) current gains and unilateral power gain (full triangles) vs. collector current density for a $2 \times 10 \mu\text{m}$ emitter device. The measured dc (full circles) and small signal (full squares) current gains at $V_{CE} = -3.5 \text{ V}$ are also shown.

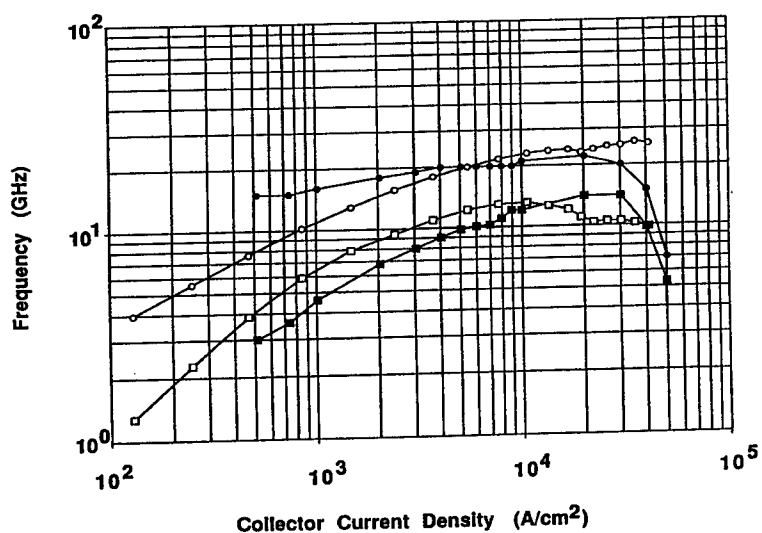


Figure 3. Calculated cutoff frequency f_t (open squares) and maximum oscillation frequency f_{max} (open circles) vs. collector current density for a $2 \times 10 \mu\text{m}$ emitter device. The measured f_t (full squares) and f_{max} (full circles) for $V_{CE} = -3.5 \text{ V}$ are also shown.

quasi-Fermi level splitting at the abrupt emitter-base junction.¹⁸ We have shown that hole drift-diffusion across the emitter-base space-charge region is of comparable importance to thermionic emission in controlling hole injection into the base. Optimization of the transistor's multilayer structure was performed, indicating that a maximum f_T as high as 23 GHz and f_{max} as high as 34 GHz (without base grading) could be obtained. Base grading was shown to improve the device's cutoff frequency by as much as a factor 1.5. While quite useful in providing guidelines for the design and growth of epitaxial multilayer structures for HBT fabrication and in assessing factors limiting device performance, the drift-diffusion approach is known to be limited in its accuracy, especially for the study of abrupt HBTs. Quantum mechanical tunneling of holes through the valence band spike at the emitter-base junction is a possibility. In fact, due to the coupling between holes in the heavy-, light-, and split-off (SO) bands, there is a finite probability of hole conversion while tunneling across an interface.

Recently, we have used the 6X6 Luttinger-Kohn Hamiltonian to study the effects of the SO band on the transmission and reflection coefficients of holes through abrupt heterointerfaces.¹⁹ For the case of potential steps, the effects of the SO band on hole tunneling coefficients were found to be quite drastic in the case of an InP/InGaAs potential step. This effect is due to the low values of the threshold energy for free propagation in the SO band on either side of the heterointerface in the InP/In_{0.53}Ga_{0.47}As system, as shown in Fig. 4. For a heavy hole incident from the left on the structure shown in Fig. 4, the tunneling probabilities (T_{HH} for heavy-to-heavy and T_{LH} for heavy-to-light hole conversion) are shown in Fig. 5.

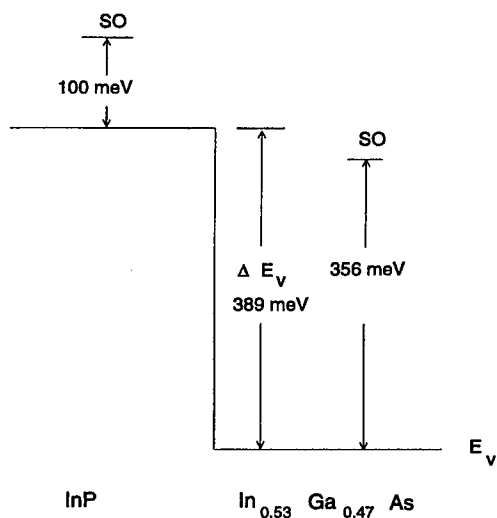


Figure 4. Diagram of the valence band discontinuity across the InP/In_{0.53}Ga_{0.47}As interface. The horizontal lines labeled "SO" are the locations of the spin-orbit split-off energy band minimum on both sides of the structure. Energy is measured positively going into the valence band.

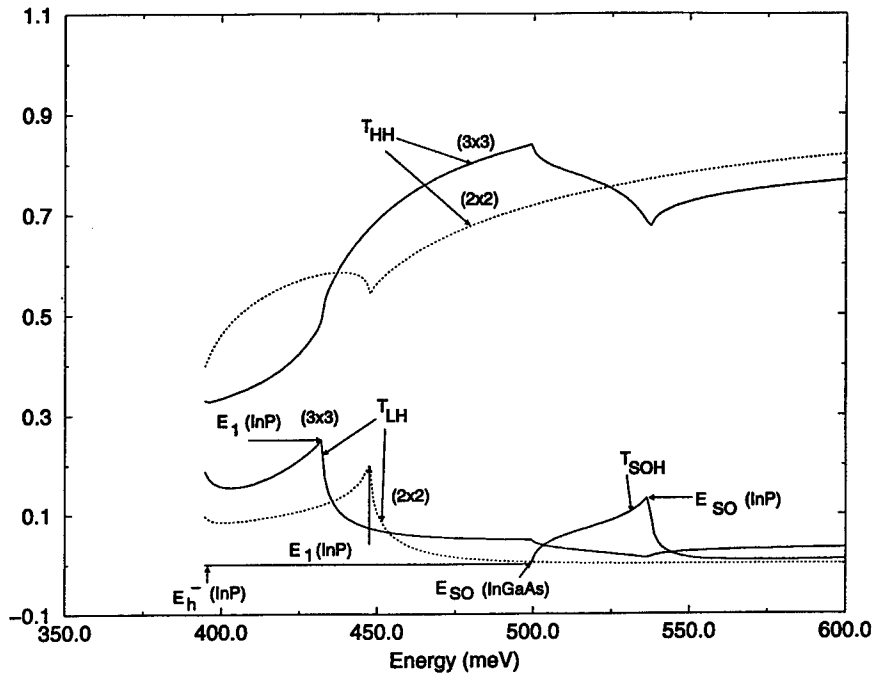


Figure 5. Transmission coefficients for an incident heavy hole at the InP/In_{0.53}Ga_{0.47}As interface shown in Fig. 4 including (curves labeled as 3x3) and ignoring (2x2) the SO band. The in-plane wavevector of the incident hole is set equal to $0.04(\pi/a)$ (where a is 5.83 Å, the InP lattice constant).

Figure 5 clearly shows that the SO band cannot be neglected in the calculation of the emitter injection efficiency of *pn*p HBTs using InP/In_{0.53}Ga_{0.47}As materials for the emitter-base junction. Because of the probability of hole conversion at the interface (especially heavy to light), the SO band also affects the energy distribution of the heavy holes injected in the base of the *pn*p HBTs. The latter controls the HBT base transit time, which is one of the leading components affecting the high-frequency performance of these devices. The problem of heavy-to-light hole conversion at the emitter-base junction of *pn*p HBTs therefore needs further study.

4. Conclusions

In this paper, we have given a brief survey of the already demonstrated wide variety of applications of state-of-the-art *pn*p HBTs. We have stressed the importance of developing accurate modeling tools to help in the design of these structures, which usually require sophisticated growth techniques. Optimization of the transistor multilayer structure has been studied,¹⁷ indicating that a maximum cutoff frequency

as high as 23 GHz and a maximum frequency of oscillation as high as 34 GHz (without base grading) can be obtained. Base grading can improve the device's cutoff frequency by as much as a factor 1.5.¹⁸

Much more work is still needed to assess the ultimate high-frequency performance of *pn*p HBTs. This endeavor is quite worthwhile, considering the huge prospect for these devices including telecommunications, microwave and millimeter-wave circuits, data converters, high speed digital circuits and optical telecommunications²⁰. We believe that the potential market of *pn*p HBTs will get a big boost from recently developed wafer fusion and bonding techniques.^{21,22}

5. Acknowledgments

This work was supported by the National Science Foundation (ECS-9525942). We also acknowledge the Ohio-Cray supercomputing center for the use of their facilities.

References

1. W. E. Stanchina, R. A. Metzger, M. W. Pierce, *et al.*, "Monolithic fabrication of *n*p*n* and *pn*p InAlAs/InGaAs HBTs," in: *Proc. 5th Intern. Conf. InP Related Mater.* (1993), p. 569.
2. W. E. Stanchina, R. A. Metzger, D. B. Rensch, *et al.*, "InP-based technology for monolithic multiple-devices, multiple-function ICs," *GOMAC-91 Dig.* (1991), p. 385.
3. A. Nakagawa and K. Inoue, "Symmetric *pn*p InAlAs/InGaAs double HBTs fabricated with Si-ion Implantation," *IEEE Trans. Electron Dev.* **13**, 285 (1992).
4. L. M. Lunardi, S. Chandrasekhar, and R. A. Hamm, "High-speed, high-current-gain *pn*p InP/InGaAs HBTs," *IEEE Electron Dev. Lett.* **14**, 19 (1993).
5. H. Fukano, Y. Kawamura, and Y. Takanashi, "High-speed InAlAs/InGaAs *pn*p HBTs," *IEEE Electron Dev. Lett.* **9**, 312 (1988).
6. C. K. Peng, T. Won, C. W. Litton, and H. Morkoç, "A high-performance InGaAs/InAlAs double HBT with nonalloyed n^+ InAs cap layer on InP(n) grown by molecular beam epitaxy," *IEEE Electron Dev. Lett.* **9**, 331 (1988).
7. W. L. Chen, J. P. Sun, G. I. Haddad, *et al.*, "InGaAs/InP hot electron transistors grown by chemical beam epitaxy," *Appl. Phys. Lett.* **61**, 189 (1992).
8. J. D. Cressler, J. Warnock, D. L. Harnage, *et al.*, "A high-speed complementary silicon bipolar technology with 12 fJ power-delay product," *IEEE Electron Dev. Lett.* **14**, 523 (1993).
9. M. E. Kim, B. Bayraktaroglu, and A. Gupta, Chapter 5 in: F. Ali and A. Gupta, eds., *HEMTs and HBTs: Device, Fabrication and Circuits*, Boston: Artech House, 1991.

10. P. M. Enquist, D. B. Slater, and J. W. Stuart, "Complementary AlGaAs/GaAs HBT $F^2L(CH^2L)$ technology," *IEEE Electron Dev. Lett.* **13**, 180 (1992).
11. D. G. Hill, W. S. Lee, T. Ma, and J. S. Harris, "AlGaAs/InGaAs strained-base *pn*p HBTs," *IEEE Electron Dev. Lett.* **11**, 425 (1990).
12. W. Liu, D. Hill, D. Costa, and J. S. Harris, "High-performance microwave AlGaAs/InGaAs *pn*p HBTs with high current gain," *IEEE Microw. Guided Wave Lett.* **2**, 331 (1992).
13. J. A. Hutchby, "High-performance *pn*p AlGaAs/GaAs HBTs: a theoretical study," *IEEE Electron Dev. Lett.* **7**, 108 (1986).
14. D. A. Sunderland and P. D. Dapkus, "Optimizing *npn* and *pn*p HBTs for speed," *IEEE Trans. Electron Dev.* **34**, 367 (1987).
15. J. S. Yuan, "PNP HBT design," *Solid State Electron.* **34**, 1347 (1991).
16. S. Shi, K. P. Roenker, T. Kumar, M. Cahay and W. E. Stanchina, "Simulation of *pn*p InAlAs/InGaAs HBTs," *IEEE Trans. Electron Dev.* **43**, 1466 (1996).
17. S. Datta, S. Shi, K. P. Roenker, T. Kumar, and M. Cahay, "Simulation and design of InP-based *pn*p HBTs," *IEEE Trans. Electron Dev.* **45**, 1634 (1998).
18. S. Datta, K. P. Roenker, and M. Cahay "A thermionic-emission diffusion model for graded-base *pn*p HBTs," *J. Appl. Phys.* **83**, 8036 (1998).
19. S. Ekbote, M. Cahay, and K. Roenker, "Influence of the spin-orbit split-off band on the tunneling of holes through heterostructures," *Phys. Rev. B* (1998).
20. L. T. Tran, J. C. Cowles, L. W. Yang, *et al.*, "Manufacturable InP-based HBT technology for low-voltage millimeter wave and microwave communications," in: *Proc. 1997 Int. Conf. InP Related Mater.*, Cape Code, MA (1997), p. 133.
21. Z.-H. Zhu, F. E. Ejeckam, Y. Qian, *et al.*, "Wafer bonding technology and its applications in optoelectronic devices and materials," *IEEE J. Selected Topics Quantum Electron.* **3**, 927 (1997).
22. A. Black, A. R. Hawkins, N. M. Margalit, *et al.*, "Wafer fusion: materials issues and device results," *IEEE J. Selected Topics Quantum Electron.* **3**, 943 (1997).

Resonance Phase Amplification — A Concept for Operating Si-Based Devices at mm Wave Frequencies?

H. Jorke, J. Weller, and J.-F. Luy

Daimler-Benz Research Center Ulm, Wilhelm-Runge-Str. 11, 89081 Ulm, Germany

1. Introduction

Since the invention of the transistor in 1947 there has been continuous progress in semiconductor technology that implied a steady reduction of the size of microwave devices and of their critical dimensions. This development was driven by the demand for higher cut-off frequencies f_T and, related to that, decreasing transit times.

The straightforward way of raising cut-off frequencies by reducing critical dimensions is finally limited by the appearance of new transport phenomena rather than by limitations of semiconductor technology, which has already arrived at precise control of vertical dimensions in the nanometer range. These phenomena are coming to light at structural dimensions that are comparable to or below the mean free path of carriers, throwing ballistic transport aspects into relief. On the other hand, phenomena related to the quantum mechanical nature of carriers are also becoming of increasing importance. While these phenomena are an obstacle that retards progress in conventional device development, they may be considered as well as a prospect to think of new high frequency device concepts that profit by these phenomena.

Commonly, the frequency where the common emitter current gain of microwave bipolar transistors becomes unity (its cut-off frequency), is related to the emitter-to-collector transit time τ_{EC} by

$$f_T = 1/2\pi\tau_{EC}, \quad (1)$$

where for the intrinsic transistor τ_{EC} is composed of the delay times of the emitter, base and collector

$$\tau_{EC} = \tau_E + \tau_B + \tau_C. \quad (2)$$

The overall phase of the collector current is composed of the injection phase angle $\phi = \phi_E + \phi_B$ (with the emitter and base transit angle $\phi_E = \omega\tau_E$ and $\phi_B = \omega\tau_B$, respectively), and the drift delay $\theta = \omega\tau_C$ in the base-collector junction. At frequencies below the cut-off frequency, the phase delay of the collector current is less than $\phi + \theta = 57^\circ$, as

$$\phi + \theta = \phi_E + \phi_B + \theta = 2\pi f(\tau_E + \tau_B + \tau_C) = 2\pi f/2\pi f_T < 1. \quad (3)$$

In common transistor operation the phase of the collector current increases with decreasing current gain, taking a maximum value around 1 (57°) at frequencies close to the cut-off frequency. However, at further increased phase delay, the output resistance of the transistor may become negative and result in renewed active behavior at higher frequencies. Such renewed active behavior was previously discussed on the basis of transit time effects. This article proposes a new mechanism for the appropriate phase delay based essentially on quantum-well injection. Once the enhanced phase delay makes the output resistance R_{22} negative, gain resonances appear at frequencies well in excess of the cut-off frequency f_T . We compare the results obtained from various implementations of resonance phase amplification — as we have called this renewed active behavior — and discuss how one obtains the necessary collector current phase shifts.

2. Coherent base transport

Previous work on the subject of operating transistors at frequencies in excess of the cut-off frequency was based on the transit time effect,¹⁻⁵ which exists in any finite sized device.

The effect describes the persistence of a current in the external circuit during the transit of a carrier pulse in the active region of the device. Including transit time effects, the output resistance of a bipolar transistor becomes:³

$$R_{22} = \text{Re}(Z_{22}) = \frac{\cos \phi - \cos(\phi + \theta)}{\omega \theta C_{BC}} |\alpha_B| + R_E \quad (4)$$

where $\phi = \phi_E + \phi_B = \omega(\tau_E + \tau_B)$ is the injection phase delay, $\theta = \omega W_C / v_s$ is the drift delay in the base-collector junction, C_{BC} the base-collector junction capacitance, $|\alpha_B|$ the base transport factor, and $R_E = kT/qI_C$ the differential emitter resistance. Equation 4 indicates that R_{22} becomes most negative when $\phi + \theta = 2\pi$. If this overall phase is acquired mainly at the expense of θ , i.e. $\phi \ll 1$, we get approximately

$$R_{22} \approx -\frac{\phi^2}{2\pi\omega C_{BC}} |\alpha_B| + R_E \quad (5)$$

This result implies that transistors with small transit times in the base (and, accordingly, with small base widths) are not suited for resonance phase amplification. On the other hand, taking a more optimum injection delay of $\phi = \pi$ (i.e. a thicker base), we obtain:

$$R_{22} = -\frac{2}{\pi\omega C_{BC}} |\alpha_B| + R_E \quad (6)$$

The difficulty with standard bipolar structures is that in this second, seemingly more favorable situation, the base transport factor $|\alpha_B|$ becomes small as the diffusive transport in a thicker base damps out the modulated structure of the incoming carrier distribution.⁴ A way out of this dilemma was shown by Luryi *et*

al.,^{4,5} who suggested the use of grading or step base structures to get an enhanced forward diffusion in the base, keeping $|\alpha_B|$ sufficiently high. In the limit of $|\alpha_B| = 1$ the base transport would be purely coherent. This limit, however, is probably restricted to cryogenic temperatures.³ Results from a room temperature implementation using a graded base profile in the Si/SiGe heterosystem are discussed below.

3. Quantum-well injection

In this implementation, the required phase delay to get resonance phase amplification is accomplished by the injection mechanism rather than by transit time effects. Figure 1 shows the schematic band diagram of an appropriate transistor structure.

Given sufficiently narrow separation between the base barrier layers, quantized subbands are formed in the double-barrier confined quantum well. If the emitter-base junction is forward biased, carriers tunnel resonantly through the base when the quasi-Fermi energy of the emitter equals the first subband in the quantum well, as illustrated in Fig. 1(b). Thus, the characteristic of the collector current exhibits a peaked structure shown in Fig. 2. A similar quantum-well diode injection device was previously suggested by Kesan *et al.*⁶

What is the output resistance of such a transistor? To get the right answer, we still have to consider transit-time effects, namely by the drift of carriers through the base-collector junction. Therefore, we are setting⁷

$$R_{22} = \frac{A(\omega)}{\omega \theta C_{BC}} (\cos \varphi - \cos(\varphi + \theta)) \quad (7)$$

with $A(\omega)$ being the injection ratio

$$A(\omega) = \frac{1}{\sqrt{1 + \omega^2 \epsilon^2 \epsilon_0^2 / \sigma^2}} \quad (8)$$

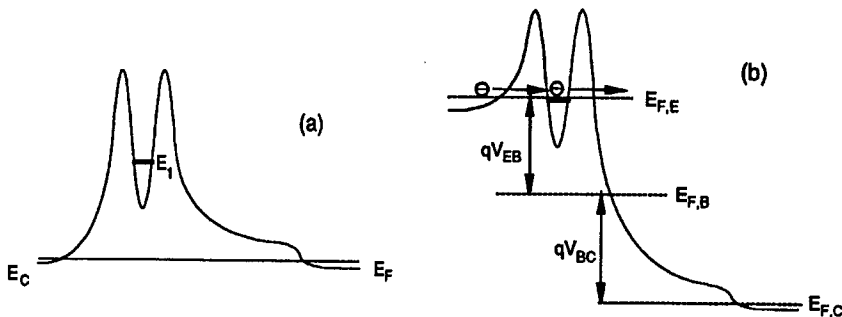


Figure 1. Schematic conduction band diagram of a double-barrier base transistor in equilibrium (a) and under active biasing conditions (b).

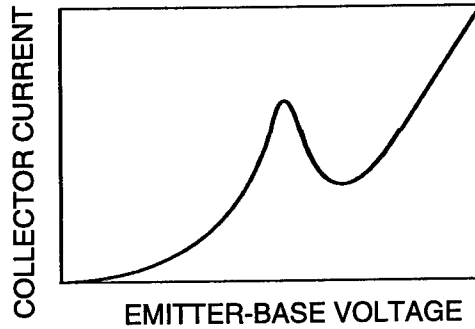


Figure 2. Collector current versus emitter-base forward bias of a double barrier base transistor where electrons are injected by resonant tunneling via a first subband in the quantum well.

and $\phi(\omega)$ the injection phase angle

$$\phi(\omega) = \tan^{-1}(\omega \epsilon \epsilon_0 / \sigma) \quad (9)$$

where σ is the injection conductivity. The injection conductivity is related to the emitter delay time τ_E according to

$$\sigma = \frac{dj_C}{dE} = \frac{1}{A} \frac{dI_C}{dE} = \frac{X_{EB}}{A} \frac{dI_C}{dV_{EB}} = \frac{\epsilon \epsilon_0}{C_{EB} R_E} = \frac{\epsilon \epsilon_0}{\tau_E} \quad (10)$$

where C_{EB} is the emitter-base capacitance and R_E the differential emitter resistance. Note that, depending on the transistor operation point, $R_E = dV_{EB}/dI_C$ is either positive or negative (Fig. 2).

Inserting Eqs. (8), (9) and (10) into Eq. (7) we finally get:

$$R_{22} = \frac{1}{\omega \theta C_{BC}} \frac{1}{1 + \omega^2 \tau_E^2} (1 - \cos \theta + \omega C_{EB} R_E \sin \theta) \quad (11)$$

$$\phi(\omega) = \tan^{-1}(\omega R_E C_{EB}). \quad (12)$$

We want to emphasize that the basic Eq. 7 (describing the quantum-well injection device) is analogous to Eq. 4 (coherent base transport device). Discrepancies arise because of the base transport factor α_B (taken to be unity in the quantum-well injection case) and the emitter transport factor α_E , that was for simplicity³ assumed to be unity in the coherent transport device (Eq. 4). The injection ratio appearing in Eq. 7 can be identified with the emitter transport factor $|\alpha_E|$ of the quantum-well injection device. It also becomes unity when $\omega \tau_E \ll 1$, as in the coherent transistor case.

The injection phase angle $\phi(\omega)$ of Eq. 12 increases linearly with ω until $|\omega \tau_E| \approx 1$. Then, depending on the operating point, it shows either a capacitive phase shift $\phi = \pi/2$ ($R_E > 0$) or an inductive phase shift $\phi = -\pi/2$ ($R_E < 0$). In the former case

carriers are injected in-phase, whereas in the latter case carriers are injected counter-phase ($\phi = 3\pi/2$).

In the following section we discuss the implementation of this device in the Si/SiGe material system and compare simulations of coherent base transport and quantum-well injection.

4. Implementation in the Si/SiGe system

In order to get concrete results that also allow a comparison of various resonance phase amplification concepts we have investigated their implementation in the Si/SiGe heterosystem. Selection of that material system is motivated largely by its compatibility with silicon technology.

To study the concept that is based on coherent base transport we have considered a series of Si/SiGe heterojunction bipolar transistors (HBTs) with compositionally graded base layers. The base grading and thickness were chosen to keep constant the base transit time τ_B , given by⁸

$$\tau_B = \frac{W_B^2}{\mu_n \Delta E_g} \left[1 - \frac{kT}{\Delta E_g} \left(1 - e^{-\frac{\Delta E_g}{kT}} \right) \right] \quad (13)$$

The minority carrier mobility μ_n is a key parameter for the design of these structures. We assumed a value $\mu_n = 400 \text{ cm}^2/\text{V}\cdot\text{s}$,⁹ which corresponds, according to Einstein's relation, to a minority diffusion constant of $D_n = 10 \text{ cm}^2/\text{s}$ at room temperature. To get an estimate of the output resistance, we still need to know the base transport factor α_B . For a graded band-gap HBT, Luryi *et al.*⁵ have derived the following expression

$$\alpha_B(\omega) = \frac{e^r}{\cosh(\lambda) + (1 + 2i\omega\tau_{B,\text{drift}}/r)^{-1/2} \sinh(\lambda)} \quad (14)$$

where $\lambda = (r^2 + 2i\omega\tau_{B,\text{diff}})^{1/2}$ and r is the ratio of the diffusion time $\tau_{B,\text{diff}} = W_B^2/2D_n$ to the drift time $\tau_{B,\text{drift}} = W_B^2/(\mu_n \Delta E_g)$. Figures 3 and 4 show the calculated gain $|U|$ (solid lines) and output resistance R_{22} (dashed lines) as a function of the frequency for $\Delta E_g = 0$ (flat base) and $\Delta E_g = 185 \text{ meV}$, respectively.

The unilateral gain U for the intrinsic transistor assumes the following form:³

$$U = \frac{|\alpha_B \alpha_C|^2}{4\omega^2 C_{BC}^2 R_B R_{22}} \quad (15)$$

with α_C being the collector transport factor and R_B the base resistance. In the flat base situation, the unloaded intrinsic transistor of Fig. 3(a) has two zeros in $R_{22}(\omega)$ corresponding to resonances in $|U|$ (solid line) at about $f = 100$ and 200 GHz . To be more realistic, in Fig. 3(b) we have added an external resistance of 15Ω to the output resistance. This addition causes the zeros in R_{22} to disappear. Correspondingly, there is no active behavior ($|U| < 1$) above the cut-off frequency.

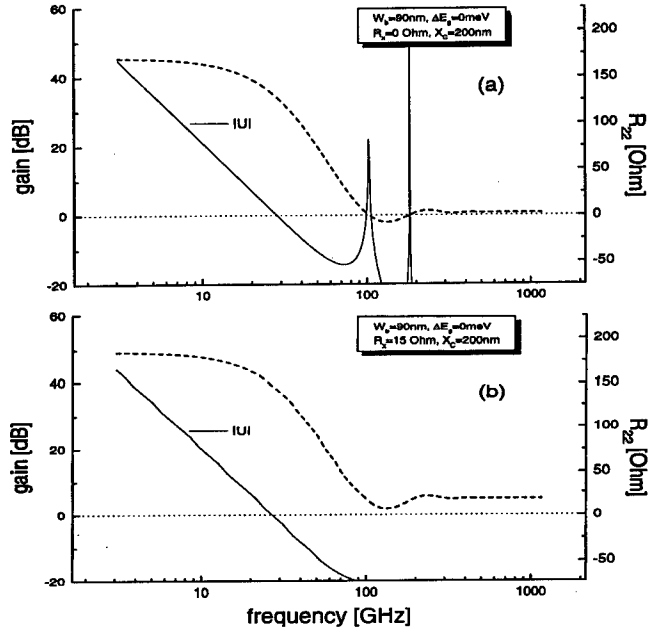


Figure 3. $|U|$ and R_z of a flat base HBT with $W_B = 90$ nm, $R_B = 50$ Ω, $C_{BC} = 30$ fF, $\tau_c = 2$ ps. The external resistance is set to zero (a) and 15 Ω (b).

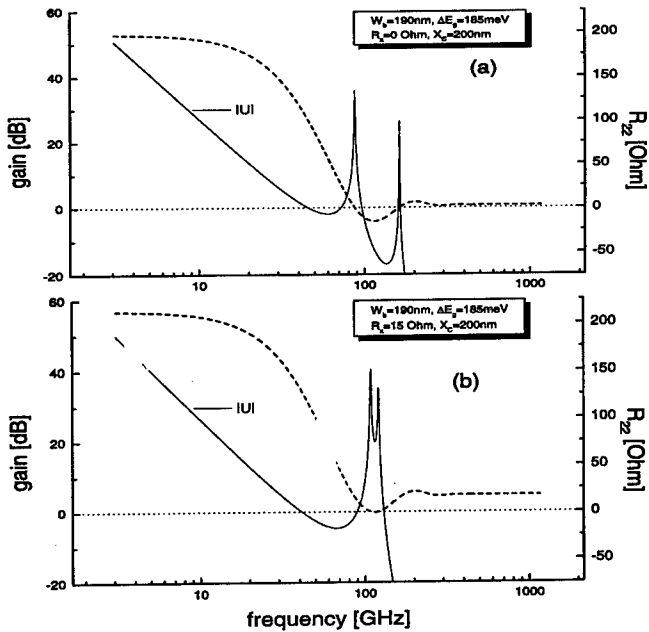


Figure 4. $|U|$ and R_z of a flat base HBT with $W_B = 190$ nm, $R_B = 24$ Ω, $C_{BC} = 30$ fF, $\tau_c = 2$ ps. The external resistance is set to zero (a) and 15 Ω (b).

Figure 4 shows results from the compositionally graded base structure ($\Delta E_g = 185$ meV). The drift of minority carriers⁵ significantly increases the maximum negative output resistance compared to the flat base situation. An external resistance of 15Ω causes the negative output impedance to decrease as well — see Fig. 4(b) — but resonances in $|U|$ are preserved.

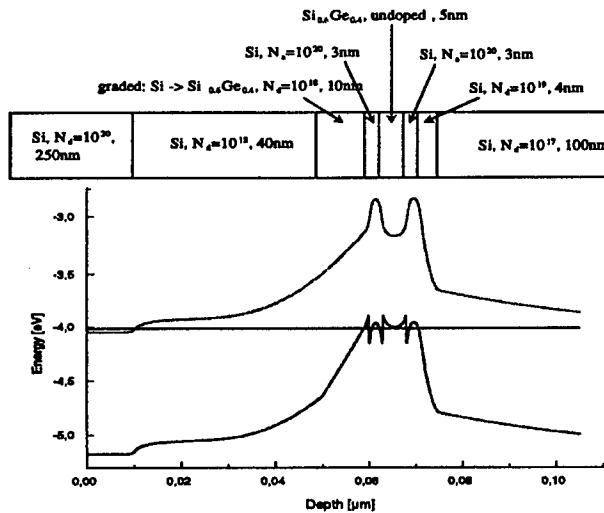


Figure 5. Layer set-up of a quantum-well injection transistor using a Si_{0.4}Ge_{0.6} layer in-between heavily *p*-doped Si layers and its energy band diagram.

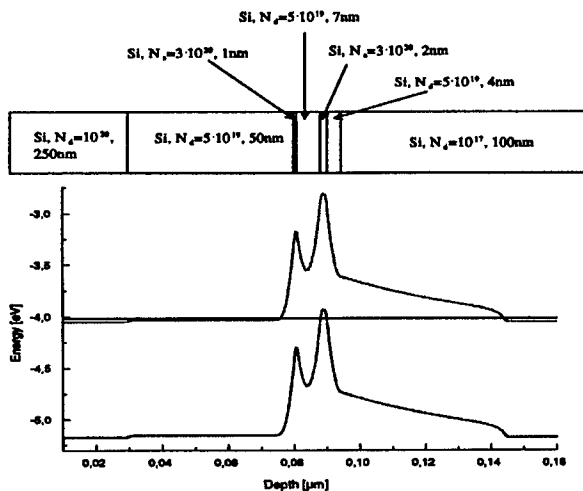


Figure 6. Layer set-up of a quantum-well injection transistor using a *p-n-p* doping sequence in the base and its energy band diagram. The barriers are designed to be symmetric under forward bias.

In our inquiry on resonance phase amplification by quantum well injection we have considered two types of structures, shown in Figs. 5 and 6 along with their band diagrams. In Fig. 5 the quantum well is formed by a $\text{Si}_{0.4}\text{Ge}_{0.6}$ layer in-between heavily p -doped Si layers. In Fig. 6 the quantum well is formed by an appropriate p - n - p doping layer sequence. An advantage of the latter is the feasibility to adjust barrier heights almost up to the band-gap of approximately 1.1 eV. On the other hand, due to a probably higher scattering rate in the well of the heavily doped p - n - p structure, resonant tunneling may be more seriously disturbed there.

A key parameter for the output resistance of the quantum well injection device is the emitter resistance, which is not known precisely. In our modeling based on Eq. 13 we have set $R_E = \pm 50 \Omega$ for in-phase and counter-phase injection respectively. Similar values were reported in a recent microwave study on resonant tunneling diodes.¹⁰ Calculated output resistance and unilateral gain are depicted in Figures 7 and 8.

For in-phase injection, shown in Fig. 7, the output resistance is qualitatively similar to the case of coherent base transport (Figs. 3, 4) but with much lower values of maximum negative resistance — a few Ω , instead of the 15 Ω in the coherent device. This result stems from the small injection angle of the quantum-well device (see Eq. 5). With respect to high-frequency operation, the in-phase quantum-well injection device behaves much like an HBT with a very thin base. Hence, the addition of a realistic parasitic resistance of 15 Ω to the output resistance R_{22} removes all resonances, as can be seen from a comparison of Figs. 7(a) and 7(b).

The situation changes for counter-phase injection, shown in Fig. 8. Without any parasitics, the unilateral gain in Fig. 8(a) is quite similar to the in-phase injection results of Fig. 7(a): in both cases, the output resistance shows multiple jumps above cut-off. In the more realistic situation that includes a 15 Ω parasitic resistance, however, only the counter-phase injection mode is still able to produce a resonance in the unilateral gain due to its considerably higher $|R_{22}|$. Where the in-phase injection mode $|R_{22}(\omega)|$ reaches values of only a few Ω , the counter-phase injection mode yields, according to Eq. 6

$$R_{22} = (\tau_c/2 - |R_E|C_{BE})/C_{BC}. \quad (17)$$

With some reasonable assumptions ($\tau_c = 4$ ps, $C_{BC} = 10$ fF, $R_E = -50 \Omega$, $C_{BE} = 100$ fF) the impedance level is found to be

$$|R_{22}| \approx 300 \Omega \quad (18)$$

which is much larger than can be obtained by in-phase injection and even larger than the impedance provided by coherent base transport devices, simulated in Figs. 3 and 4.

So, quantum-well injection operated in a counter-phase mode appears to be a promising concept for producing a resonance in the unilateral gain at elevated frequencies. Due to the high impedance level associated with that kind of device changes in the sign of the output resistance can be readily obtained.

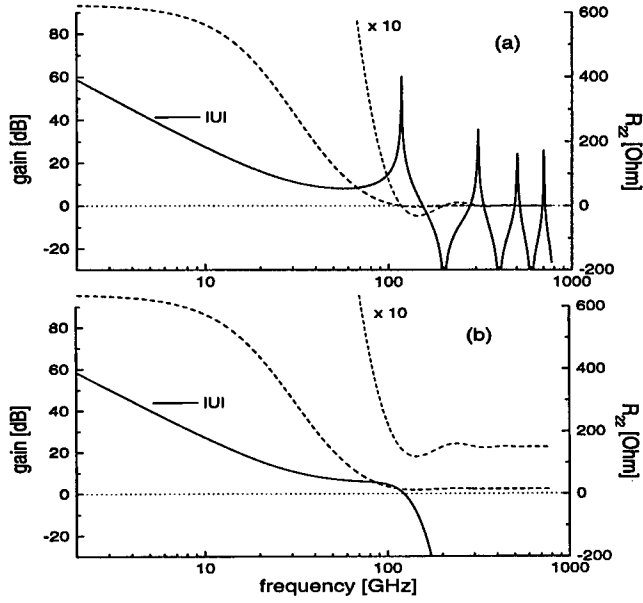


Figure 7. $|U|$ and R_{z2} of a quantum-well transistor operated in an in-phase injection mode with $R_E = 50 \Omega$, $C_{EB} = 100$ fF, $R_B = 50 \Omega$, $C_{BC} = 12$ fF, $X_C = 500$ nm, and $\tau_c = 5$ ps. The external resistance is set to zero (a) and 15Ω (b).

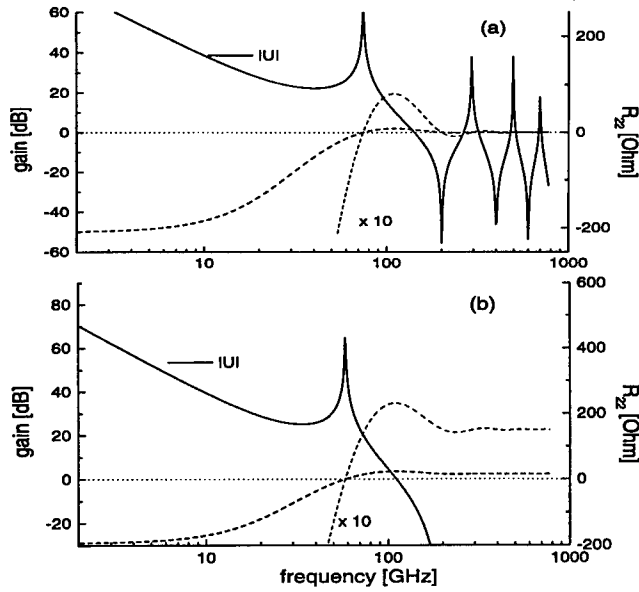


Figure 8. $|U|$ and R_{z2} of a quantum-well transistor operated in an in-phase injection mode with $R_E = -50 \Omega$, $C_{EB} = 100$ fF, $R_B = 50 \Omega$, $C_{BC} = 12$ fF, $X_C = 500$ nm, and $\tau_c = 5$ ps. The external resistance is set to zero (a) and 15Ω (b).

5. Conclusions

We compare various ways of increasing the injection phase delay ϕ and drift delay θ in bipolar transistors to get a negative output resistance R_{22} and thereby renewed active behavior at frequencies above the cut-off. In conventional bipolar transistors — where "conventional" refers to an ungraded "flat base" — increasing ϕ is insufficient because of a concomitant decay in the base transport factor α_B . This detrimental decay is weakened in grading or step base concepts that profit by forward enhanced diffusion of carriers in the base. The use of quantum-well injection is a totally different approach to enhance ϕ that produces a favorable injection phase delay of $\phi = 3\pi/2$.

6. Acknowledgments

The work was supported by the Bundesministerium für Bildung und Forschung under grant 01 M 2959 A. The authors alone are responsible for the contents.

References

1. G. T. Wright, "Small-signal theory of the transistor transit-time oscillator (translator)," *Solid State Electron.* **22**, 399 (1979).
2. S. Tiwari, "Frequency dependence of the unilateral gain in bipolar transistors," *IEEE Electron Dev. Lett.* **10**, 574 (1989).
3. A. A. Grinberg and S. Luryi, "Coherent transistor," *IEEE Trans. Electron Dev.* **40**, 1512 (1993).
4. S. Luryi, "Ultrafast operation of heterostructure bipolar transistors resulting from coherent base transport of minority carriers," *Proc. 1993 Dev. Res. Conf.*, Charlottesville, VA (1993), p. 59.
5. S. Luryi, A. A. Grinberg, and V. B. Gorfinkel, "Heterostructure bipolar transistor with enhanced forward diffusion of minority carriers," *Appl. Phys. Lett.* **63**, 1537 (1993).
6. V. P. Kesan, D. P. Neikirk, B. G. Streetman, and P. A. Blakeley, "A new transit-time device using quantum-well injection," *IEEE Electron Dev. Lett.* **8**, 129 (1987).
7. J.-F. Luy, "Transit-time devices," in: J.-F. Luy and P. Russer, eds., *Silicon-Based Millimeter Wave Devices*, Berlin: Springer-Verlag, 1994, p. 49.
8. H. Krömer, "Two integral relations pertaining to the electron transport through a bipolar transistor with a nonuniform energy gap in the base region," *Solid State Electron.* **28**, 1103 (1985).
9. J. Weller, H. Jorke, K. Strohm, *et al.*, "Assessment of transport parameters for the design of high speed Si/SiGe HBTs with compositionally graded base," to appear in *Thin Solid Films* (1998).
10. G. Cohen and D. Ritter, "Microwave performance of $\text{Ga}_x\text{In}_{1-x}\text{P}/\text{Ga}_{0.47}\text{In}_{0.53}\text{As}$ resonant tunneling diodes," *Electronics Lett.* **34**, 1267 (1998).

Silicon Quantum Integrated Circuits

D. J. Paul

Cavendish Laboratory, University of Cambridge, Madingley Road, Cambridge, CB3 0HE, U.K.

B. Coonan, G. Redmond, G. M. Crean

National Microelectronics Research Centre, Lee Maltings, Prospect Row, Cork, Ireland

B. Holländer, S. Mantl

Institut für Schicht-und-Ionentechnik, Forschungszentrum Jülich, 52425 Jülich, Germany

I. Zozoulenko, K.-F. Berggren

Dept. of Physics, University of Linköping, Linköping, S-58183, Sweden

J.-L. Lazzari, F. Arnaud D'Avitaya, and J. Derrien

CRMC2-CNRS, Campus de Luminy, Marseille, 13288 Cedex 9, France

1. Introduction

While III-V devices have consistently demonstrated superior performance to silicon, silicon is still the dominant technological semiconductor with III-V technology having a minute percentage of total sales. One may put forward many subtle and non-subtle reasons to account for this situation, but there is one that dominates all — *cost*. The low cost of CMOS may be traced to the fabrication of billions of nearly identical transistors on wafers of ever-increasing diameter. The fabrication processes and the device performance rely heavily on silicon's physical and chemical properties, as well as silicon-compatible insulators. SiO_2 and Si_3N_4 have allowed Si to dominate over faster materials such as GaAs, which requires more expensive fabrication schemes that cannot reach the phenomenal yields and hence integration densities achievable on CMOS production lines.

With the amount of capital and knowledge presently tied up in Si production and research, the impetus to design and produce Si devices is enormous. With production plants costing in excess of \$1 billion, it is almost impossible to persuade companies to change to completely new, untried technologies. CMOS is so cheap and dominant that one must find applications where it cannot be used, such as optoelectronics, analog, or high-speed (e.g. rf) if a new technology is to appear in the marketplace. One compromise is to use a new CMOS-compatible material system, such as $\text{Si}_{1-x}\text{Ge}_x$,^{1,2} to allow bandgap-engineered devices with higher performance or new functionality. Increased integration levels reduce the

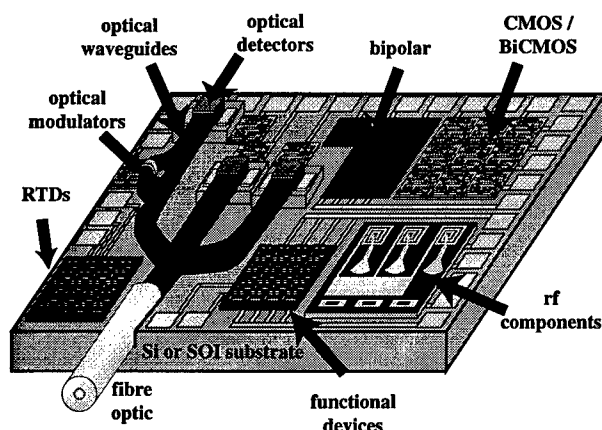


Figure 1. A schematic of the system-on-a-chip of the future.

number of chips in final products and hence should lower costs and increase yields. The dream is to produce complete system-on-a-chip single-chip solutions, illustrated in Fig. 1. The integration of dynamic random access memory (DRAM) onto logic chips has also demonstrated the difficulty of integrating different optimized fabrication routes, with yields dropping as both memory and logic are combined. Multi-chip modules (MCMs) may therefore be used to combine two or three chips that may have individually optimized fabrication processes and easier individual testing. The simplest type of device one may integrate with CMOS is the heterojunction bipolar transistor (HBT), where a narrow $\text{Si}_{1-x}\text{Ge}_x$ ($x < 0.3$) base layer is inserted into a BiCMOS fabrication process. HBTs with f_T up to 120 GHz³ and f_{max} up to 150 GHz⁴ have been demonstrated from research lines, although BiCMOS production HBTs show more modest values of 50–60 GHz.⁵ The important issue in designing such wafers is to incorporate the $\text{Si}_{1-x}\text{Ge}_x$ base with the least amount of changes possible to a normal CMOS or BiCMOS process.⁶ Impressive yields have been demonstrated from 200 mm diameter wafer production lines⁶ and numerous applications have appeared in the literature.⁷

2. The heterostructure field effect transistor

While the HBT is now a production device for analog and rf applications, a Si-based FET type device in these areas would potentially have enormous impact. GaAs MESFETs and high electron mobility transistors (HEMTs) have led the way in this field. The mobile telephone is the application now driving the market, although many other portable devices are predicted to expand the low power and high frequency applications.

Initial experiments in the field concentrated on pseudomorphic SiGe layers grown on a Si substrate. One of the limiting factors in the performance of CMOS is the p -MOS device, which has a mobility about 2.5 times lower than the n -MOS device. To compensate, the transistors have to be scaled accordingly to balance the current drive. Initial attempts to improve the p -MOS used a pseudomorphic $\text{Si}_{1-x}\text{Ge}_x$ ($x < 0.3$) channel to improve the performance. The best results using CMOS processing led to only about 20% improvement,⁸ so it remains cheaper to scale to smaller gate-lengths than include SiGe in the CMOS process.

Some of the most exciting SiGe FET results to date are from modulation-doped FETs (MODFETs). Figure 2 plots f_T against gate length L for a number of different device families.⁹⁻¹³ The best SiGe devices have f_T comparable to GaAs HEMTs and better than GaAs MESFETs for constant L .¹⁰ P -type devices have shown even more impressive performance with f_T up to 70 GHz at $L = 0.1 \mu\text{m}$.¹³ Mobilities at 300 K (77 K) of 2830 (18000) $\text{cm}^2/\text{V}\cdot\text{s}$ at 2×10^{12} (8×10^{11}) cm^{-2} for the n -MODFET¹⁴ and 1300 (14000) $\text{cm}^2/\text{V}\cdot\text{s}$ at 1.5×10^{12} (1.0×10^{12}) cm^{-2} for the p -MODFET¹⁵ have been demonstrated, with even higher mobilities at cryogenic temperatures (4.2 K and below).^{16,17} Transconductances in the $L = 0.25 \mu\text{m}$ n -MODFET were 330 (600) mS/mm at 300 K (77 K),¹⁰ while the $L = 0.1 \mu\text{m}$ p -MODFET reached 237 mS/mm .¹³ The n -channel enhancements are due to strain lifting the valley degeneracy and reducing intervalley scattering, thereby allowing the saturation velocity to be achieved at lower electric fields (Fig. 3). The p -channel enhancements are due to the lower effective mass and Ge-like strain-modified valence band structure. Initial modeling of circuit performance is also encouraging: n -MODFETs with loads of 200 fF were shown to exhibit 560 ps delays in NOR gates at 1.1 V, compared to 1400 ps delays for the equivalent CMOS.¹² The CMOS had to be run at 3.3 V to achieve the same delay, consuming nine times the power of the SiGe MODFETs. The major problem with MODFETs in regard to CMOS-compatible processing is the enhanced dopant diffusion even at temperatures as low as 650°C .¹⁸

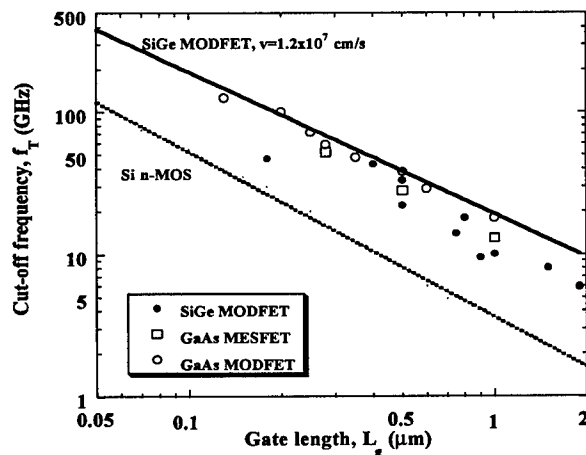


Figure 2. Cut-off frequency f_T vs. gate length L for different device families.⁹⁻¹³

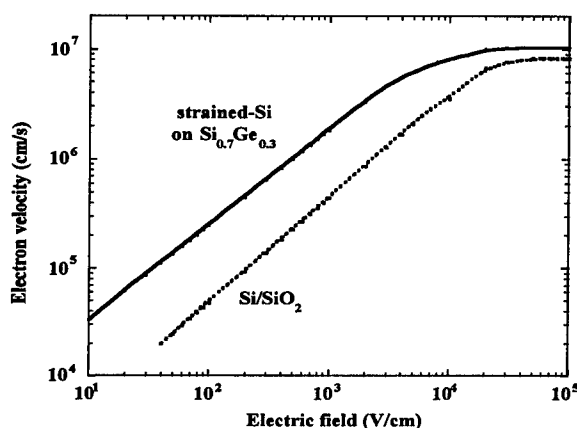


Figure 3. Electron velocity vs. electric field in MOSFETs and Si/SiGe MODFETs.¹⁰

Numerous theoretical calculations have shown that the valley splitting in strained Si produces higher electron velocities for a given electric field or alternatively for low power applications, a given electron velocity may be produced for a reduced electric field (Figure 3).^{10,19} One choice of structure would therefore be a strained-Si MOSFET. For the present research, a SiGe cap is inserted between the oxide and the strained-Si channel to form a HFET. While the addition of the cap will reduce the transconductance due to the increased gate to channel spacing, the HFET is predicted to have lower noise characteristics than a strained-Si MOSFET.²⁰ Considering the conservative nature of the semiconductor industry, this choice allows a transistor that is similar to a normal MOSFET in operation but is predicted to have substantially enhanced performance. This choice of structure also allows all the major fabrication stages from a CMOS line to be characterized and modeled with respect to the new material. This permits a full set of process simulation tools to be developed, allowing more complicated SiGe devices to be produced later and also allows the easy transfer of the technology to CMOS production lines.

3. Material

The material for the HFETs has been bought from a commercial supplier, SiGe Microsystems of Ottawa, Canada. The growth system was an ultra-high vacuum CVD system originally designed by IBM and manufactured by Leybold.²¹ The heterostructure design consisted of a 1 μm thick linearly graded buffer (up to Si_{0.75}Ge_{0.25}), a 1 μm thick Si_{0.75}Ge_{0.25} buffer, a 10 nm thick Si quantum well, a 20 nm Si_{0.75}Ge_{0.25} spacer and a 4 nm thick Si cap. Future batches will have far more aggressive scaling of layers, once any major limitations from the first batch have been discovered and a full working set of process simulation tools is available to optimize the structure.

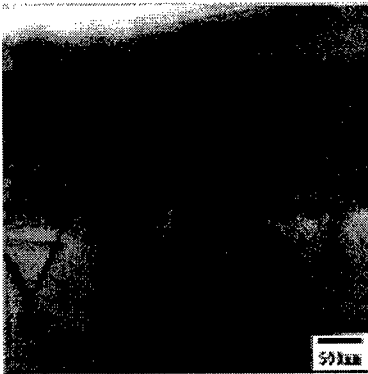


Figure 4. XTEM of HFET material.

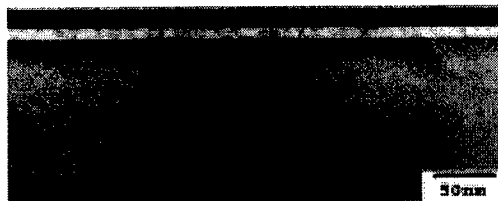


Figure 5. XTEM of Si quantum well.

Figure 4 shows a cross sectional transmission electron micrograph (XTEM) of the HFET wafer structure. The bottom layers contain many threading dislocations in the graded part of the buffer with the top layers relatively dislocation free as is typical with such virtual substrates.²² Several dislocations protruding from the buffer into the substrate were found. Very few dislocations were found to run from the graded buffer region to the constant Ge content buffer. Figure 5 exhibits a higher resolution XTEM of the top layers with the strained Si quantum well clearly visible about 20 nm below the surface. The layers are extremely flat and smooth on the scale of the micrograph, although optical micrographs in Nomarski contrast display the typical characteristic crosshatch pattern observed in all virtual substrates.²²

4. Fabrication process

A standard 1.5 μm ground rule CMOS process is being used in the present research to fabricate the HFET. Gate lengths from 100 μm to a sub-ground rule of 0.5 μm are being tested using a standard poly-Si self-aligned process and a second low resistivity, non self-aligned Al gate process. There are a number of processes which require slight modifications when Si/Si_{1-x}Ge_x heterostructure material is added to the CMOS process. It is important, however, to try to minimize the number of changes to allow easy transfer of a process to manufacture. For successful device fabrication a number of key steps must be achieved: device isolation, gate oxide formation, polysilicon (Al) gate formation, dopant activation, source/drain contact formation, device passivation and metallization. Below, a number of the key issues will be discussed.

The major advantage CMOS processing has over every other semiconductor device family is the use of a high quality native oxide that may be used both as an insulator and as a highly selective etch stop. The major problem in this approach for strained heterostructures is that SiO₂ is normally grown at 900 °C or higher temperatures. At such temperatures, Ge diffusion, dopant diffusion and strain relaxation may all occur. Any processing scheme for the HFET must therefore use a reduced thermal budget compared to a standard CMOS process. As device

geometries reduce in size in CMOS lines, the thermal budget must be reduced also to prevent doping diffusion and so SiGe HFETs and CMOS will become thermally more compatible in the future.

In *n*-type MODFETs enhanced dopant diffusion has been demonstrated at temperatures as low as 600 °C,¹⁸ probably due to enhanced diffusion through defects or the competing segregation effects of As and Ge as observed during SiGe growth.²³ On the other hand, it has been shown that 30 minute anneals at 800 °C and 3 minute rapid thermal annealing (RTA) at 950 °C do not reduce the 77 K mobility of undoped HFETs.²⁴ The removal of modulation doping from the heterostructure will therefore not only remove the remote ionized impurity scattering but also improve the thermal stability. The philosophy of using an HFET design over a MODFET may be justified therefore from potential device performance, fabrication issues and CMOS compatibility. Most CMOS processes use a number of high-temperature anneals for implant activation, thermal oxide growth and dopant diffusion in poly-Si gates. In the present work, all the required anneals have been condensed into one anneal to substantially reduce the total thermal budget.

A secondary problem related to any SiO₂ growth may also be identified. Any growth of a thermal oxide on Si_{1-x}Ge_x layers has demonstrated a Ge pile up at the SiO₂(GeO₂)/Si_{1-x}Ge_x interface, resulting in a high interface state density.²⁵ One solution is to oxidize a thick Si cap layer, but then a second strained Si quantum well may form between the oxide and the Si cap if the oxide does not consume enough Si.²⁶ This second well produces a parasitic parallel conducting layer that significantly reduces transistor performance. Low temperature plasma-enhanced grown oxides have also been fabricated that demonstrate acceptable properties on Si_{1-x}Ge_x layers^{27,28} and Si_{0.7}Ge_{0.3} MODFETs.²⁹

Deposited oxides are another possibility. In our devices, we measured midgap interface state densities as low as $2 \times 10^{11} \text{ cm}^{-2} \text{ eV}^{-1}$ for oxides CVD-deposited at 300 °C on relaxed Si_{0.77}Ge_{0.23} wafers. This interface state density produces a flat band voltage of -1.8 V, which compares favourably to -0.965 V for a thermally grown oxide on Si of similar thickness.

For device isolation, a trench etch is required. A significant amount of research on the reactive ion etching (RIE) of Si_{1-x}Ge_x material has appeared in the literature.^{30,31} Although the chemistry changes with the addition of Ge, it is not difficult to produce the required etching. Device passivation is accomplished using a low temperature plasma enhanced CVD SiO₂ passivation layer. Since this layer is deposited at 300 °C, there is little impact on the total thermal budget for the HFET fabrication. Metallization is carried out using a standard CMOS scheme with 0.5 μm thick Al (1% Si) being used. A forming gas anneal at 400 °C for 1 hour is the final step to alloy the metal without causing spiking of the Al.

5. Quantum devices

While journals have been publishing numerous papers annually on quantum devices for a number of decades, only a few simple III-V tunnel diodes have

appeared in the marketplace for ultra-high speed applications. The problem with many designs is that while the devices may switch at incredible speeds and/or use low power, the devices cannot communicate with conventional electronics or operate at room temperature. Many quantum devices require asynchronous architectures and slow substantially when placed in synchronous circuits. For quantum devices to be useful, the devices must be able to communicate with CMOS circuits for input/output and connection to the world as this is the cheapest solution. Operation at 77 K may be useful for a few specialized high-cost applications, but for the mainstream electronics markets and especially for the growing portables market, devices must be able to operate at room temperature.

The first RTD was produced in 1974 in GaAs/AlGaAs³² but only in the last few years have useful circuits appeared in the literature. The performance has steadily increased with improved material quality. From a circuit point of view, RTDs allow low-power memories and high-speed logic where the high performance does not involve the aggressive lithographic scaling required of CMOS.³³ Thus, 50 nW SRAM cells have been demonstrated³⁴ that are 200 times lower in power consumption than conventional 0.1 μm SRAMs. While the ultimate speeds for these low-power devices are slower than many competitors, access times less than 1 ns have been predicted at large levels of integration. Generic logic circuits operating at over 12 GHz for 20 to 48 μm^2 lithographic feature sizes³⁵ have also been demonstrated experimentally.

An ideal RTD in the Si system would use SiO_2 barriers with standard Si processing. For useful tunneling currents, however, this scheme requires oxides of

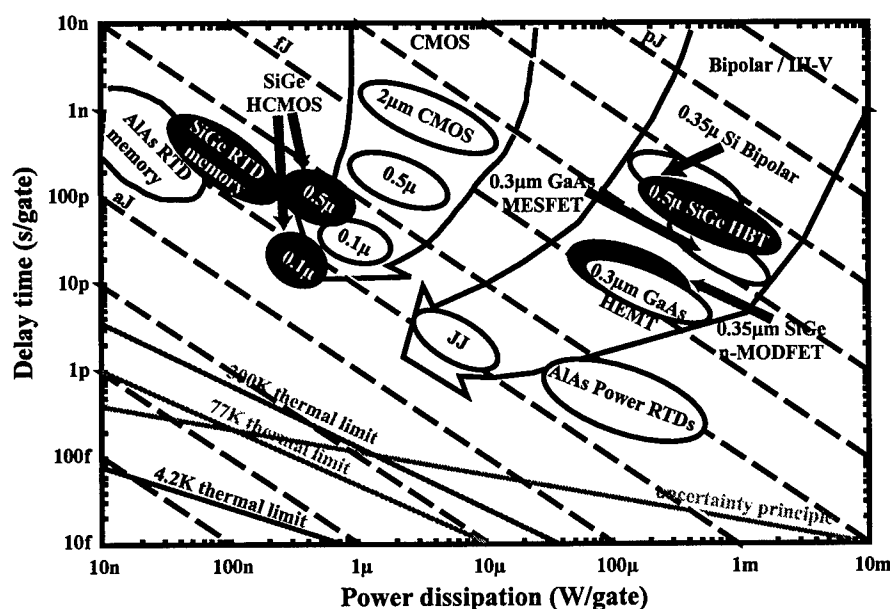


Figure 6. Power-delay performance of competing devices (After Refs. 5, 10-13, 34, 35, 37).

about 1 to 2 nm thickness without any pinholes or defects. Such an oxide or CMOS compatible insulator is presently beyond the capability of technology and so most Si-based RTDs have used Si/SiGe heterolayers.³⁶ It is possible to epitaxially grow a *p*-type RTD directly on Si (100) but since the strained SiGe layers are quantum wells and the relaxed Si layers are barriers, designs of such devices rely on relatively small valence band discontinuities and can only operate at low temperatures. The *n*-type RTD has strained Si wells and relaxed SiGe barriers, and the larger conduction band discontinuities might allow $T = 300$ K operation despite the use of a virtual substrate. For a high-quality Si-based RTD to operate at room temperature, reduction of the thermionic emission must be achieved.

6. Conclusions

The methodology, potential performance, fabrication issues and applications of Si based HFETs have been described. The main issues required to produce a manufacturable product involve the production of a device with significantly higher performance than an equivalent MOSFET and a fabrication process which is as close to a normal CMOS fabrication process as possible. We have introduced CMOS-compatible processing of resonant tunneling diodes and discussed some of the fabrication problems. Finally, Figure 6 plots the predicted power-delay performance of the devices considered in the present work against the major device families of CMOS, bipolar and III-V metal semiconductor FETs (MESFETs) and MODFETs. The HFETs and RTDs have potential applications in low power applications and memory respectively.

7. Acknowledgments

The work in this paper is funded by the European Community under the Esprit Microelectronics Advanced Research Initiative (MEL-ARI), Project No. 22987.

References

1. D. J. Paul, "Silicon germanium heterostructures in electronics: the present and the future," *Thin Solid Films* **321**, 172 (1998).
2. F. Schäffler, "High-mobility Si and Ge structures," *Semicond. Sci. Technol.* **12**, 1515 (1997).
3. E. F. Crabbé, B. S. Meyerson, J. M. C. Stork, and D. L. Harame, "Vertical profile optimization of very high frequency epitaxial Si-and SiGe-base bipolar transistors," *IEDM Tech. Digest* (1993), p 83.
4. A. Gruhle and A. Schuppen, "Recent advances with SiGe heterojunction bipolar transistors," *Thin Solid Films* **294**, 246 (1997).

5. D. L. Hareme, J. H. Comfort, J. D. Cressler, *et al.*, "Si/SiGe epitaxial-base transistors," *IEEE Trans. Electron Dev.* **42**, 455 (1995).
6. D. A. Sunderland, D. C. Ahlgren, M. M. Gilbert, *et al.*, "Manufacturability and applications of SiGe HBT technology," *Solid State Electronics* **41**, 1503 (1997).
7. J. D. Cressler, "SiGe HBT technology: a new contender for Si-based rf and microwave circuit applications," *IEEE Trans. Microw. Theory Techn.* **46**, 571 (1998).
8. S. Verdonckt-Vanderbroek, E. F. Crabbé, B. S. Meyerson, *et al.*, "SiGe-channel heterojunction *p*-MOSFETs," *IEEE Trans. Electron Dev.* **41**, 90 (1994).
9. S. M. Sze, ed., *High-Speed Semiconductor Devices*, New York: Wiley, 1990.
10. K. Ismail, "Si/SiGe high speed field-effect transistors" *IEDM Tech. Digest* (1995), p 509.
11. R. Hagelauer, T. Ostermann, U. König, M. Glück, and G. Höck, "Performance estimation of Si/SiGe hetero-CMOS circuits," *Electronics Lett.* **33**, 208 (1997).
12. U. König, M. Glück, A. Gruhle, *et al.*, "Design rules for *n*-type SiGe hetero-FETs," *Solid State Electronics* **41**, 1541 (1997).
13. M. Arafa, K. Ismail, J. O. Chu, B. S. Meyerson, and I. Adesida, "A 70 GHz f_T low operating bias self-aligned *p*-type SiGe MODFET," *IEEE Electron Dev. Lett.* **17**, 586 (1996).
14. K. Ismail, S. F. Nelson, J. O. Chu and B. S. Meyerson, "Electron transport properties of Si/SiGe heterostructures: Measurements and device implications," *Appl. Phys. Lett.* **63**, 660 (1993).
15. U. König and F. Schäffler, "P-type Ge-channel MODFETs with high transconductance grown on Si substrates," *IEEE Electron Dev. Lett.* **14**, 205 (1993).
16. K. Ismail, M. Arafa, K. L. Saenger, J. Chu, and B. S. Meyerson, "Extremely high electron mobility in Si/SiGe modulation-doped heterostructures," *Appl. Phys. Lett.* **66**, 1077 (1995).
17. Y. H. Xie, D. Monroe, E. A. Fitzgerald, *et al.*, "Very high mobility two-dimensional hole gas in Si/Ge_xSi_{1-x}/Ge structures grown by molecular beam epitaxy," *Appl. Phys. Lett.* **63**, 2263 (1993).
18. D. J. Paul, J. M. Ryan, P. V. Kelly, *et al.*, "Investigations of electron-beam and optical induced damage in high mobility SiGe heterostructures," *Solid State Electronics* **41**, 1509 (1997).
19. Th. Vogelsang and K. R. Hoffmann, "Electron transport in strained Si layers on Si_{1-x}Ge_x substrates," *Appl. Phys. Lett.* **63**, 186 (1993).
20. A. G. O'Neill and D. A. Antoniadis, "Investigation of Si/SiGe-based FET geometries for high frequency performance by computer simulation," *IEEE Trans. Electron Dev.* **44**, 80 (1996).
21. H. Lafontaine, D. C. Houghton, D. Elliot, *et al.*, "Characterisation of Si_{1-x}Ge_x epilayers grown using a commercially available ultrahigh vacuum chemical vapour deposition," *J. Vac. Sci. Technol. B* **14**, 1675 (1996).

22. F. K. LeGoues, B. S. Meyerson, and J. Morar, "Anomalous strain relaxation in SiGe thin films and superlattices," *Phys. Rev. Lett.* **66**, 2903 (1991).
23. S. M. Hu, D. C. Ahlgren, P. A. Ronsheim, and J. O. Chu, "Experimental study of diffusion and segregation in a Si-(Ge_xSi_{1-x}) heterostructures," *Phys. Rev. Lett.* **67**, 1450 (1991).
24. H. Klauk, T. N. Jackson, S. F. Nelson, and J. O. Chu, "Thermal stability of undoped strained Si channel SiGe heterostructures," *Appl. Phys. Lett.* **68**, 1975 (1996).
25. D. K. Nayak, J. S. Park, J. C. Woo, K. L. Wang, and I. C. Ivanov, "Interface properties of thin oxides grown on strained Ge_xSi_{1-x} layer," *J. Appl. Phys.* **76**, 982 (1994).
26. U. König and F. Schäffler, "Si/SiGe modulation doped field effect transistor with two electron channels," *Electronics Lett.* **27**, 1405 (1991).
27. P. W. Li, H. K. Liou, E. S. Yang, *et al.*, "Formation of stoichiometric SiGe oxide by electron cyclotron resonance plasma" *Appl. Phys. Lett.* **60**, 3265 (1992).
28. I. S. Goh, J. F. Zhang, S. Hall, W. Eccleston, and W. Kerner, "Electrical properties of plasma-grown oxide on MBE-grown SiGe," *Semicond. Sci. Technol.* **10**, 818 (1995).
29. N. Griffin, D. J. Paul, M. Pepper, *et al.*, "Gating high mobility silicon germanium heterostructures," *Microelectronic Eng.* **35**, 309 (1997).
30. G. S. Oehrlein, G. M. W. Kroesen, E. Defresart, Y. Zhang, and T. D. Bestwick, "Studies of the reactive ion etching of SiGe alloys," *J. Vac. Sci. Technol. A* **9**, 768 (1991).
31. Y. Zhang, G. S. Oehrlein, E. Defresart, and J. W. Corbett, "Reactive ion etching of SiGe alloys using fluorine containing plasmas," *J. Vac. Sci. Technol. A* **11**, 2492 (1993).
32. L. L. Chang, L. Esaki, and R. Tsu, "Resonant tunnelling in semiconductor double barriers," *Appl. Phys. Lett.* **24**, 593 (1974).
33. S. M. Sze, ed., *Modern Semiconductor Device Physics*, New York: Wiley, 1998, pp. 306-317.
34. J. P. A. van der Wagt, A. C. Seabaugh and E. Beam III, "RTD/HFET low standby power SRAM gain cell," *IEDM Tech. Digest* **96**, 425 (1996).
35. W. Williamson III, S. B. Enquist, D. H. Chow, *et al.*, "12 GHz clocked operation of ultralow power interband resonant tunnelling diode pipelined logic gates," *IEEE J. Solid State Circuits* **32**, 222 (1997).
36. H. C. Liu, D. Landheer, M. Buchanan, and D. C. Houghton, "Resonant tunnelling in Si/Si_{1-x}Ge_x double barrier structures," *Appl. Phys. Lett.* **52**, 1809 (1988).
37. A. W. Wieder, "Si-microelectronics: perspectives, risks, opportunities, challenges," in: S. Luryi, J. M. Xu, and A. Zaslavsky, eds., *Future Trends in Microelectronics*, NATO ASI Series Vol. 322, Dordrecht: Kluwer, 1996, pp. 13-21.

RSFQ Computing: The Quest for Petaflops

M. Dorojevets

Dept. of Electrical and Computer Engineering, SUNY at Stony Brook, Stony Brook, NY 11794

P. Bunyk, D. Zinoviev, and K. K. Likharev

Dept. of Physics and Astronomy, SUNY at Stony Brook, Stony Brook, NY 11794

1. Introduction

After several years of decline, high performance computing is again high on the U.S. national agenda. A recent letter from the President's Informational Technology Advisory Committee names high-end computing as one of three key areas to be supported by the federal government.¹ In particular, it recommends establishing "the goal of attaining sustained petaops/petaflops [performance] on real applications" by the end of the next decade. A long list of important tasks for the petaflops computing includes nuclear stockpile stewardship, fluid dynamics modeling for aerospace system development, chemical reaction simulation for new drug design, climate and ocean modeling for longer-term weather forecasts, and global economy dynamics modeling.

Let us see what would it take to reach this ambitious goal using CMOS technology which will be available by the middle of the next decade. According to the most authoritative industrial forecast,² by the year 2006 high-performance microprocessors may reach a clock frequency of 2 to 3.5 GHz, and feature up to 200 million transistors on $\sim 5 \text{ cm}^2$ chips dissipating the power up to 160 W each. Comparing these numbers to those for the present-day advanced microprocessors such as Pentium II or Alpha 21264, the peak performance of such a future multiprocessor chip can be crudely estimated as 10 Gflops. Hence, in order to achieve the peak performance of 1 petaflops will take approximately 100K chips like this, with the total power dissipation of the order of 15 MW. The management of power of such proportions would take a sizeable building rather than just a large hall.

We would like to stress that this discouraging estimate stems from a very *optimistic* assumption of 70-nm fabrication technology, for which there are "no known solutions".² Moreover, to sustain a performance close to this peak level, such a system would require medium/coarse grain parallelism at the level of 10^5 independent instruction streams in a user's program. (We assume that the parallelism within each instruction stream will be exploited at the single-chip level). Significant (300-ns scale) latency of interprocessor communication in a system of such a physical size is also a negative factor, prone to stalling processors

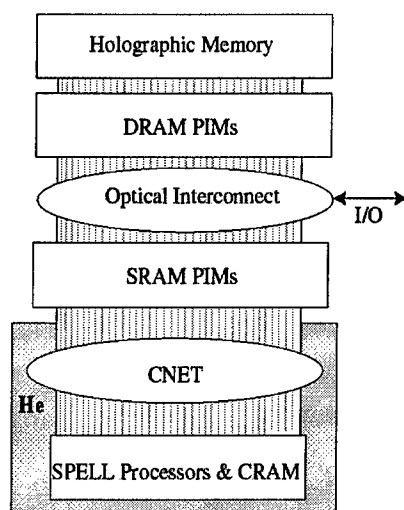


Figure 1. Hybrid technology multithreaded architecture system concept.

if interprocessor communication and synchronization is a large enough fraction of the computational process.

These problems associated with semiconductor processors have stimulated a search for alternative approaches to petaflops-scale computing. In particular, we are participating in the hybrid technology multithreaded architecture (HTMT) project^{3, 4} led by Jet Propulsion Laboratory. The goal of this project is to carry out a preliminary study of a computer architecture and the functional organization of a system that would utilize novel electronic and optoelectronic technologies to achieve petaflops-level performance by or soon after the year 2005. The HTMT system has a hierarchical organization (see Fig. 1) with multiple levels of distributed memory: holographic data storage (HRAM), semiconductor SRAM and DRAM, and cryomemory (CRAM), as well as three types of processors: SRAM- and DRAM-based processors-in-memory (PIMs) operating at room temperature, and RSFQ superconductor processing elements (SPELLs) operating at liquid helium temperature.

RSFQ (Rapid Single-Flux-Quantum) family of ultrafast superconductor logic circuits has been developed during the last decade.^{5,6} Physically, these circuits consist of superconductor loops closed with overdamped Josephson junctions. These loops may store and process digital bits in the form of single quanta of magnetic flux ($\Phi_0 = h/2e \approx 2 \times 10^{-15}$ Wb), and can transmit and receive single-bit signal in the form of short (picosecond) voltage pulses $V(t)$ with quantized area $\int V(t)dt = \Phi_0 \approx 2$ mV-ps. The HTMT project places RSFQ logic at the heart of the HTMT system as its main means of number crunching in SPELLs, as well as switching and routing in an interprocessor communication network (CNET), because of its two unique features:

- very high speed (simple RSFQ devices have been demonstrated at frequencies up to 750 GHz,⁷ approximately 15 times faster than any semiconductor devices of comparable complexity), and
- extremely low power consumption (of the order of 10^{-18} J/bit, i.e. some five orders of magnitude less than that in advanced CMOS circuits).

In addition to these advantages, RSFQ circuit fabrication technology using low temperature superconductors (mostly, niobium) is considerably simpler than CMOS, especially since it does not require deep-submicron patterning.⁵

Unfortunately, the necessity of deep refrigeration of RSFQ circuits does not allow this exciting new technology to compete with CMOS for most digital electronic applications. However, in a high performance system such as a petaflops computer the refrigeration costs would be a negligible component of the total cost, and the advantages of RSFQ shine bright. As will be shown below, this technology may allow us to reach the peak performance of 1 petaflops using a very compact RSFQ core with 4K SPELLs, occupying physical space as small as 0.2 m³ and dissipating power as low as 1 kW.

The design of the HTMT system, and especially its RSFQ subsystem, is far from trivial. The time-of-flight of a signal over 1 cm distance on an RSFQ chip is about 100 ps. This is why these circuits, operating with a-few-ps clock cycle, are essentially relativistic: two gates on the same chip may be well outside of each other's light cone. More importantly, at such speed, unavoidable picosecond-scale clock skew makes global timing impractical. This makes us re-think microprocessor design techniques. Another problem stemming from the unparalleled speed of the RSFQ logic is how to hide the enormous CRAM/SRAM latency visible to SPELL processors in the HTMT system. This latency varies from ~70 processor cycles when accessing the local CRAM to several hundreds of cycles when accessing SRAM. The DRAM level is not visible to SPELL processors.

A more general formulation of these questions is whether the picosecond RSFQ components can be integrated with other relatively slow components in a way that would allow the whole system to achieve the sustained petaflops level performance on important tasks, while keeping cost, power consumption, and software and programming complexity much lower than those for traditional, silicon-based approaches. Answering this question involves all levels of computer design, including parallelizing compilers, program execution model, instruction set architecture, as well as processor, memory, and network structure. This is why the HTMT effort is essentially multi-disciplinary.^{3,4}

In collaboration with other participating groups, our Stony Brook team has completed a design study of the RSFQ subsystem.⁸ The objective of this article is to describe in brief our design approach and key features of the subsystem. Since the design is still evolving, all the concrete numbers given below should be considered as preliminary. Still, our article presents the architecture and organization of the superconductor rapid single-flux-quantum (RSFQ) subsystem of a planned "petaflops computer" capable of performing 10^{15} floating-point

operations per second. If actually built, such a system would be almost 1,000 times faster than the most powerful systems available in 1998. Preliminary estimates show that the 0.8 μm RSFQ technology enables the implementation of processing elements (SPELLs) operating with an average processing rate of 100 GHz, cryo-memory (CRAM) with 120-ps cycle time, and an interprocessor communication network (CNET) with the bandwidth of 30 Gbps per channel. An RSFQ system with 4,096 SPELLs will be sufficient to provide the peak performance of 1 petaflops, while occupying a physical space as small as 0.2 m^3 and dissipating power as low as 1 kW (at 4K). This implies 300-kW total power consumption at room temperature, orders of magnitude less than that for a hypothetical petaflops system implemented using any prospective semiconductor transistor technology.

2. Multithreading, HTMT program execution model, and COOL ISA

A powerful method of hiding latency is *multithreading*. This technique reduces the processor idle time by overlapping the execution of unrelated tasks called *threads*. Multithreaded architectures have been studied since the 1960s when the technique was first implemented in the peripheral processors of the CDC 6600. Several high-performance multithreaded computers, including HEP,⁹ MARS-M,¹⁰ and Tera,¹¹ have actually been built, and the concept of multithreading has been studied in several research projects.¹²⁻¹⁵

Multithreading and context prefetching have been accepted as the key techniques of latency tolerance in the HTMT execution model.¹⁶ In this model, PIMs perform pre-processing of a program to find ready threads and allocate their contexts in CRAM. When a SPELL finishes the execution of a thread, SRAM PIMs fetch the results from CRAM into SRAM and transfer them to DRAM/HRAM if necessary. All these multilevel activities (searching for ready threads, pre-allocating thread contexts in CRAM, executing threads in SPELLs, and transferring data from CRAM to SRAM) can be performed in parallel provided there is enough parallelism in programs.

The HTMT model exposes and exploits two types of parallelism:

- coarse-grain parallelism represented by threads (which are essentially parallel function/procedure/process invocations),
- medium-grain parallelism represented by program entities called *strands* inside threads (e.g., parallel loop iterations).

This two-level multithreading manifests itself in COOL instruction set architecture developed for RSFQ processors.¹⁷ COOL is a parallel 64-bit RISC architecture with the support for two-level simultaneous multithreading and pseudo-vector computation. The latter is instrumented with so-called *quad* instructions, each of which is able to perform operations on short 4-word vectors. In contrast to truly vector architectures, COOL ISA does not rely on any vector registers.

Input/output operands of the quad operations are to be fetched from or placed into four adjacent data registers.

The goal of utilizing in full the advantages of RSFQ circuits was the driving force behind our decision to include these architectural concepts in the COOL ISA.

3. RSFQ circuit design principles

The main peculiarity of RSFQ circuits from the point of view of computer design is that most Boolean logic functions (NOT, AND, etc.) are performed *by latching gates*. Functionally, a latching gate may be considered as an indivisible combination of a logic gate and an output latch. Depending on the design tasks, this feature may be considered either as a blessing or a handicap. On the positive side, it gives latches for free. For a single-cycle pipeline, however, this feature limits the number of logic levels per pipeline stage to 1 (compared to 5-10 levels in a typical CMOS design) if only latching gates are used.

Fortunately, two more elementary components are available in RSFQ. Some of the logic functions (for example, OR¹⁸) may be carried out by *non-latching circuits* which resemble traditional combinational elements. Moreover, such operations as signal fork and join are also performed by simple non-latching circuits ("splitter" and "merger", respectively⁵). Finally, an important contribution to RSFQ circuit latency may be given by passive and active *interconnects*. Passive superconductor microstrip lines can transfer RSFQ signals at any on-chip distance with speed v approaching the speed of light in vacuum ($v/c \sim 0.4$). Active Josephson transmission lines have longer delays but are able to restore RSFQ signals to their nominal amplitude.

The timing characteristics of latching gates and non-latching components, and hence their possible use for pipelining, differ significantly. A latching gate may only be occupied by one bit of data, and the next data may enter only after the gate has been cleared by a clock pulse signifying the end of the clock period. The minimum acceptable clock period T is determined by the shape of the RSFQ pulse tails, plus a necessary allowance for effects of thermal fluctuations and local variations of Josephson junction parameters. For the 0.8 μm RSFQ technology planned for an RSFQ petaflops computer, the minimum acceptable value of T , which guarantees a reasonable bit error rate and fabrication yield, is close to 10 ps.

On the other hand, non-latching components (including transmission lines) may allow multiple pulses to travel close to each other without any harmful interference. The minimum interval τ between these pulses is considerably less than T (for the 0.8 μm technology, close to 3 ps). This means that these components may be used when necessary as single-bit pipelined FIFO queues. A disadvantage of this approach is that the transmission along the non-latching components cannot be controlled from outside with any clock signal.

These RSFQ features create an unusual and substantial difference between the notions of clock cycle period and latency in instruction pipelines. A simple RISC CMOS pipeline is divided into stages which are separated by latches, so that the

signal delay (latency) L per stage equals the clock period T . An RSFQ pipeline consists of *macrostages*, each similar in function to the traditional RISC stage. Each macrostage, however, is composed of several *microstages*, each usually including one latching gate, non-latching components, and transmission lines. (We refer to this feature as *ultrapipelining*.) As a result, the latency of each macrostage can be presented as

$$L \approx n_L \times T + n_A \times \tau + a/v, \quad (1)$$

where n_L is the number of microstages, n_A the number of non-latching components in the critical path, and a the physical length of the path. Even if the two last components are negligible, the ratio L/T is at least $n_L \gg 1$. For example, recent studies of various RSFQ designs for integer carry-lookahead adders^{18,19} have shown that a 32-bit Kogge-Stone adder implemented with 0.8 μm technology has the minimum value of T between 5 and 10 ps, while L ranges between 250 and 300 ps, giving the L/T ratio between 25 and 60, depending on the timing technique employed.

This ratio represents the required number of independent instructions needed to fill out one pipeline macrostage (e.g., Integer Execute with the adder) in order to achieve its peak performance. The load/store and floating-point pipelines will need even more instructions to reach their peak rates. It is well-known, however, that this level of parallelism could be found almost exclusively in scientific programs working with regularly-structured data objects. Thus, the capability of issuing multiple instructions during loop (or vector) processing of such data objects is a critical requirement for RSFQ processors.

The solutions for 100-GHz RSFQ processors, however, cannot be the same as those for superscalar or VLIW processors (like loop unrolling, software pipelining, and etc.). The lack of a global clock makes the idea of simultaneous non-local events impractical and, as a result, the implementation of compiler-controlled, synchronous VLIW computation unrealistic. The complexity of even moderate multi-way superscalar implementation is also prohibitive.

The proposed design solution is to make each integer unit pipeline of any SPELL shareable among multiple strands, i.e. low-level instruction streams running simultaneously within each (higher-level) parallel thread. Because floating-point pipelines are longer and need more instructions to be filled out, all floating-point units of a SPELL are shared by all instruction streams of all threads running in parallel within a processor. Finally, each CRAM component is made accessible to any instruction stream running in the whole RSFQ subsystem, though we expect data traffic between SPELLs and their local CRAMs to be prevailing.

4. RSFQ technology assumptions

Though the completed stage of the RSFQ subsystem design was mostly based on preliminary, back-of-the envelope calculations rather on detailed simulations (which are still in progress), even this preliminary stage required certain

assumptions of the RSFQ technology level. So far the most complex RSFQ integrated circuits (of up to several thousands of Josephson junctions) have been built using a commercially available $3.5\text{ }\mu\text{m}$ fabrication technology, allowing a maximum clock frequency about 30 GHz. RSFQ circuits, however, obey simple scaling rules^{4,5} that have been confirmed in experiments with $1.5\text{ }\mu\text{m}$ and $0.5\text{ }\mu\text{m}$ devices. This scaling shows that VLSI circuits implemented with $0.8\text{ }\mu\text{m}$ RSFQ technology (which has been accepted as the target technology for the full-scale petaflops system) may have the following major characteristics:

- an average on-chip clock rate of 100 GHz, with power dissipation of $0.1\text{ }\mu\text{W}$ per logic gate, and
- memory cycle of 120 ps (for 512×64 bit banks), with power consumption of $0.002\text{ }\mu\text{W}$ per memory cell.

We have also assumed the integration scale up to 3 million Josephson junctions (about 300 thousand latching gates) per 4-cm^2 logic chip, and up to 10 million junctions (about 2 Mb) per memory chip of the same size. Such technology is physically possible and, given adequate investment (on the scale of \$30 million), may be available in 3 to 4 years. In the future, the transfer to mid-submicron (say, $0.4\text{ }\mu\text{m}$) RSFQ technology seems feasible, with a 50% increase in circuit performance and a dramatic (10-fold) increase in circuit density, because Josephson junctions of this size become intrinsically overdamped and bulky external shunts are no longer necessary.^{4,7}

5. SPELL processor organization

Figure 2 shows the block diagram of the proposed processor.⁸ Each SPELL has 6 arithmetic floating-point functional units operating with an average floating-point cycle time of 30 ps, altogether providing the peak performance of about 200 Gflops per SPELL. The processor also includes 16 64-bit integer functional units operating with a 10-ps cycle that can perform floating-point compare operations, giving the peak performance of 1,600 Gflops per SPELL on such operations. These 16 integer units provide the peak integer performance of 8×10^5 MIPS per SPELL (3.2×10^9 MIPS per system with 4K SPELL processors).

In order to organize the instruction streams so that all functional units can operate in parallel, each SPELL has 16 thread context units (TCU). Each TCU simultaneously issues instructions from up to 8 independent instruction streams (strands) and executes all integer, control, and floating-point compare operations. Logically, each context unit implements a register context for an active thread. Physically, each unit consists of separate issue/control logic for each of the strands and shared resources including a data register file and an integer unit. TCU can communicate with the 6 floating-point functional units, other thread units, and processor-memory interface (PMI) via the intra-processor network (PNET). Thus, the floating-point units and CRAM are shared resources for all $8\times 16 = 128$ instruction streams from 16 TCUs within each SPELL. 16 TCUs of a SPELL are

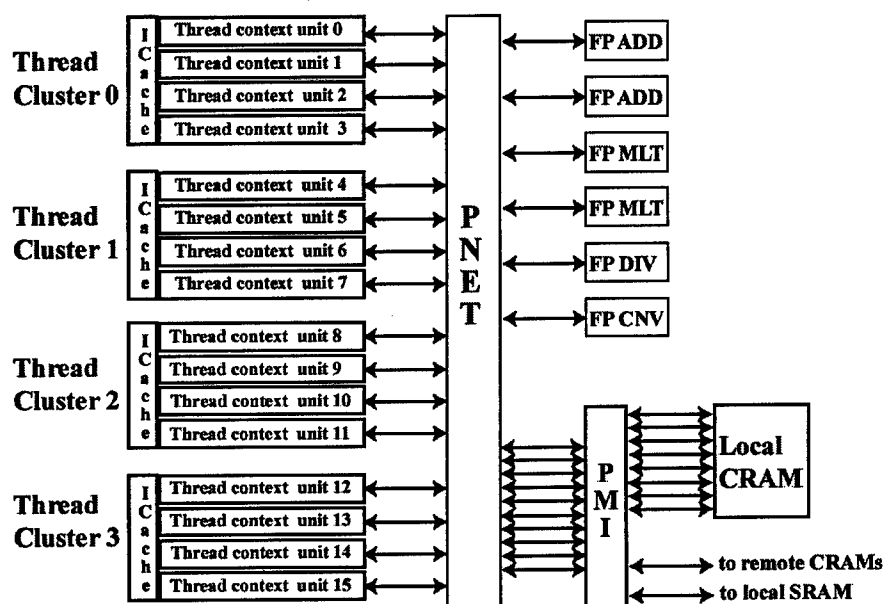


Figure 2. Top structure of the superconductor microprocessor (SPELL).

combined into 4 thread clusters, each with an 8-KB multi-port instruction cache (IC) shared by 4 TCUs within the cluster, so that each cache can service 32 instruction streams. Such shared instruction caches are beneficial when several threads work with the same code, because only one copy of the code needs to be fetched from CRAM into IC.

The PNET connects TCUs, FPUs, and PMI together using a backpressure mechanism to resolve conflicts and to buffer request/response packets in the PNET internal nodes. This mechanism prevents any packet from either leaving the network until a receiver outside PNET (e.g., a floating point unit) is able to accept it, or entering the PNET if it is congested. In the latter case, the packet waits in the TCU/FPU output buffer until PNET is able to receive the packet. PNET has 32 two-way I/O ports: 16 to thread contexts, 6 to FPUs, and 10 to the processor-memory interface, all operating at 30 Gbps.

Figure 3 shows the structure of the SPELL instruction pipeline consisting of a TCU datapath, PNET, CRAM and FPUs. As in the simple RISC processor, the instruction pipeline is divided into five stages (here, *macrostages*): Instruction Fetch, Instruction Decode/Register Fetch, Integer Execute, Memory Access/FP Execute, and Write Back. The difference with the standard CMOS design is that, first, each macrostage is divided into several microstages and, second, that up to 8 instruction streams can run simultaneously within one TCU. Except for Memory Access/FP Execute macrostage, only instructions from different instruction streams (strands) can be simultaneously processed within each macrostage. Enforcement of this rule is possible without a big loss of throughput, because the

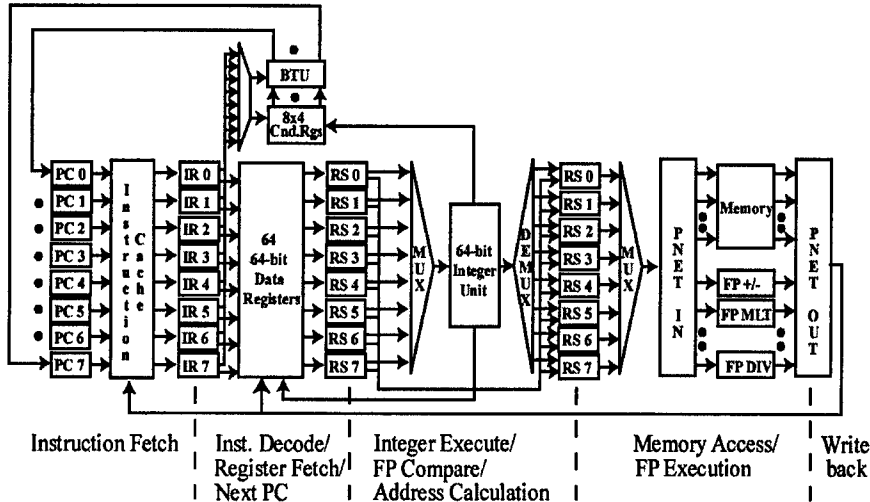


Figure 3. SPELL instruction pipeline.

time-of-flight of a synchronization signal over the length of one macrostage is considerably less than the stage latency L .

Each TCU includes a unified set of 64 data registers, each of which is able to hold either 64-bit long integer or double-precision floating-point data in IEEE 754 format. There are also $8 \times 4 = 32$ 4-bit condition registers used for branch control, and miscellaneous registers for thread control, synchronization, and exceptions handling.)

Instruction streams within each TCU share almost all the 64 data registers, the instruction cache, and the integer functional unit. Hardware reserved for each strand includes a program counter, instruction register with strand control logic, 4 condition registers, and reservation stations where operations and their input operands are placed before issuing them to integer/floating-point units and memory. The TCU components are connected by multiplexing/demultiplexing networks that are very similar to PNET in structure and functions. Messages from different streams have no compiler-assigned priorities, and a local first-come-first-served policy is used in all network switches to resolve conflicts during the message routing.

A branch/thread control unit (BTU) always begins execution of any thread in TCU from strand S0, using the initial instruction address loaded from CRAM. (It is placed into CRAM as part of a thread control block prepared by SRAM PIM.) Creation and termination of other strands is carried out using special "create/terminate strand" instructions and requires neither involvement of the runtime system nor allocation of any CRAM resources. There are no hardware limits on the number of outstanding memory references from each strand. The

only reason for an individual strand to be suspended by the hardware is the detection of a data/control hazard in the SPELL pipeline.

All types of dependencies (flow, anti- and output data dependencies) among instructions are enforced by distributed scoreboard-like logic that sets/clears a Wait bit associated with each general-purpose and special-purpose register. When an operation is issued, its destination register is marked as "not ready" by setting its Wait bit to 1. When the result is written into the register, its Wait bit is cleared (set to 0).

6. CRAM and CNET

Two other important components of the RSFQ subsystem are CRAM and CNET. Data storage in CRAM is based on the same principle as one used in RSFQ logic: each bit is presented by presence/absence of a single flux quanta in a superconductor loop closed with a Josephson junction. However, read/write operations also use another ("latching") mode, where signals are presented by voltage/current levels rather than picosecond pulses. (This mode, while not providing as high speed as RSFQ logic, is more convenient in memory cell matrices where speed is essentially limited by time-of-flight rather than by switching speed.) Generally, the memory cell structure accepted for CRAM is very close to that developed earlier by NEC.²⁰ However, in order to avoid the multi-GHz rf power supply, cell readout interferometers use overdamped Josephson junctions. The memory cells are organized in 512×64-bit matrices, with one 64-bit word occupying an entire row of this "bank". Estimates show that memory clock cycle may be close to 120 ps. Pipelined decoders are based on RSFQ logic. Together with PNET and processor-memory interface delays, the local CRAM latency as seen by streams is about 700 ps.

Each SPELL is served by 2 local CRAM chips, each with 16 banks (256 KB per chip, i.e. 512 KB per SPELL), so that the total volume of CRAM in the 4K-processor system is 2GB. Each CRAM chip has 4 two-way 30-GHz ports which can provide one-way peak bandwidth of eight 64-bit words each 30 ps, i.e., 2 TB/s per SPELL (8 PB/s in a 4K-processor system). The local CRAM/SRAM communication (through a room-temperature interface with 16K wires with 8 Gbps bandwidth) bandwidth is one word every 30 ps (i.e., 0.25 PB/s) in each direction per each local link.

The peer-to-peer and remote SRAM communications will be provided by CNET. Preliminary simulation results²¹ show that CNET, implemented as a pruned mesh of $\sim 10^5$ elementary 2×2 switches with credit-based flow control, may provide a sufficient bandwidth, while its internode, 4-layer wiring would fit on the available CMCM area of $\sim 20 \text{ m}^2$ (see below).

At this design stage when detailed system simulations have yet to be completed, we can only give very crude estimates of the expected sustained floating-point performance. Assuming the same frequencies of FP arithmetic and compare instructions (28% and 6%, respectively) as those given in Ref. 22 for five

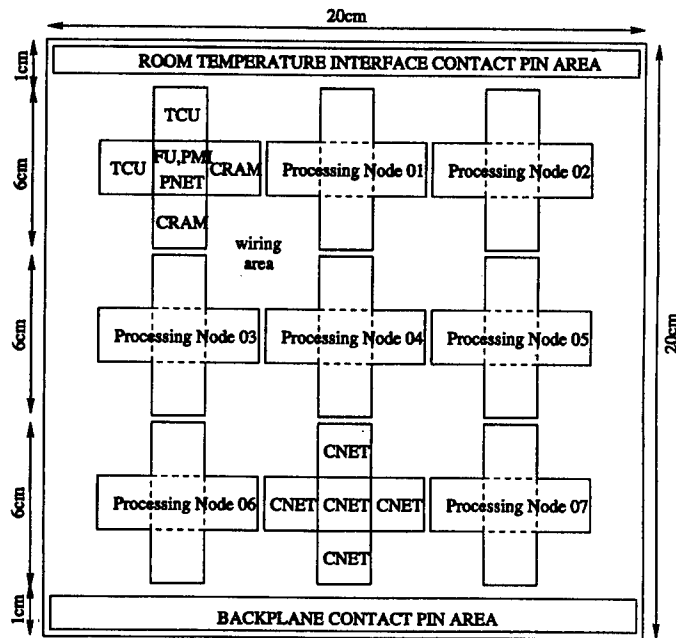


Figure 4. Planned layout of the RSFQ multi-chip module (CMCM).

programs from SPECfp92, we could calculate the maximum sustained performance (in the absence of any hazards) as 150 Gflops per SPELL, i.e. about 600 teraflops for the 4K-SPELL system.

7. What would it look like?

Physically, each processing module (SPELL with its local CRAM) of the RSFQ subsystem can be implemented as a set of five 2×2 cm² chips including:

- two double-cluster TCU chips, each with 2.8M Josephson junctions, 1,900 contact pads, and 24 mW of power dissipation (at 4K),
- one chip housing 6 floating-point functional units, processor-memory interface, and PNET, totaling 1.6M Josephson junctions, 5,800 contact pads, and 16 mW of power dissipation,
- two CRAM chips, each with 10M junctions, 3,200 contact pads, dissipating 4 mW of power.

The chips should be flip-chip mounted on a 20×20 cm cryo multi-chip module (CMCM), physically a silicon wafer with four layers of 12-μm-pitch superconducting microstrip lines. Such CMCM is large enough to house 45 chips,

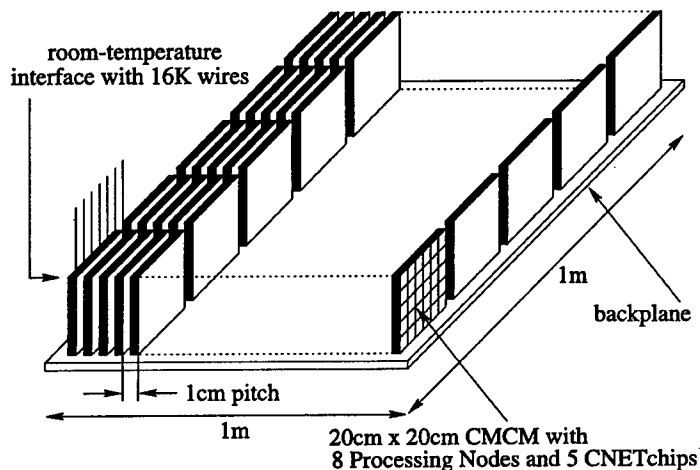


Figure 5. Possible configuration of the RSFQ subsystem.

including 8 processing modules plus 5 additional CNET chips (see Fig. 4). RSFQ circuits on each CCM would dissipate 0.5 W power, a factor of 3 less than the estimated thermal load imposed by the copper wires of the room-temperature interface. The latter number might be dramatically reduced using high temperature superconductor interconnects between the 4 K and 77 K stages of the cryostat. However, even with the current conservative estimates, an overall power load of the RSFQ subsystem at 4K is about 1 kW. For this power level, the existing efficiency of helium refrigeration of 0.3% (i.e. 20% of the perfect, Carnot efficiency) results in the room-temperature power of 0.3 MW, on the same scale as a present-day supercomputer facility with a sub-teraflops performance.

The latter power would be released in the helium liquefier/recondenser which may be remote, and does not affect CCM packaging density. Estimates show that 8 mm gaps between CCMs would be sufficient for the removal of the dissipated heat at a very moderate speed of liquid helium flow (~ 0.4 mm/s). When mounted on a backplane with such gaps (see Fig. 5), 512 CCMs necessary for the whole petaflops system will occupy a volume of 0.2 m^3 , apparently a small fraction of the total volume of the petaflops computer. This compactness of the RSFQ core would reduce the interprocessor communication latency to ~ 10 ns.

8. Conclusions

The implementation of a petaflops computer no bigger than a size of single room and with reasonable power within the next few years is an exciting opportunity. So far, our study has demonstrated that the design of the RSFQ subsystem of such a computer is feasible. We see, however, a lot of challenges, many of which still need to be addressed in more depth than has been done so far. Some of these

problems (e.g., the substantial sensitivity of RSFQ circuits to thermal fluctuations and variations of technological parameters, and the latching behavior of most logic gates) are specific for that particular technology. However, we believe that most of the new design challenges (such as the lack of a global clock) would be typical for any digital technology approaching the 100-GHz frontier; and that we are facing these problems just because RSFQ technology has arrived at this frontier first. This is why we believe that our results will provide glimpses of light on the road to the *terra incognita* of supercomputing.

9. Acknowledgements

Useful discussions with other members of the Stony Brook RSFQ System Group, especially Peter Litskevich, Yuri Pogudin, and Stas Polonsky, as well as Guang Gao, Peter Kogge, Burton Smith, Thomas Sterling, and other members of the HTMT collaboration are gratefully acknowledged. The HTMT project is supported by DARPA, NSA, and NASA. This research at Stony Brook is also supported by the NSF (ECS-9700313).

References

1. See the President's Information Technology Advisory Committee's "Interim Report to the President", August 1998, at <http://www.ccic.gov/ac/interim/>
2. "The National Technology Roadmap for Semiconductors. 1997 Edition", Semiconductor Industry Association, San Jose, CA.
3. G. Gao, K. Likharev, P. Messina, and T. Sterling, "Hybrid technology multithreaded architecture," *Proc. 6th Symp. Frontiers of Massively Parallel Computation*, Los Alamitos, CA: IEEE Comp. Soc. Press, 1996, p. 98.
4. T. Sterling, "A hybrid technology multithreaded architecture for petaflops computing," CACR, Caltech, Pasadena, CA, 1997.
5. K. Likharev and V. Semenov, "RSFQ logic/memory family: a new Josephson junction technology for sub-terahertz clock frequency digital systems," *IEEE Trans. Appl. Supercond.* 1, 3 (1991).
6. K. Likharev, "Superconductors speed up computation," *Phys. World* 10, 39 (May 1997).
7. W. Chin, V. Patel, A. Rylyakov, J. Lukens and K. Likharev, "Superconductor digital frequency dividers operation up to 750 GHz," submitted to *Appl. Phys. Lett.* (1998).
8. P. Bunyk, M. Dorofeyevs, K. Likharev, and D. Zinoviev, "RSFQ subsystem for the HTMT petaflops computing," *Technical Report 03*, RSFQ System Group, SUNY, Stony Brook, NY, 1997.
9. B. Smith, "Architecture and applications of the HEP multiprocessor computer system," in: *SPIE Real Time Signal Processing IV*, New York: SPIE, 1981, p. 241.

10. M. Dorojevets and P. Wolcott, "The El'brus-3 and MARS-M: recent advances in Russian high-performance computing," *J. Supercomputing* 6, 5 (1992).
11. "Tera: principles of operation," Seattle, WA: Tera Computer Company, 1998.
12. H. Hirata, K. Kimura, S. Nagamine, Y. Mochizuki, A. Nishimura, Y. Nakase, and T. Nishizawa, "An elementary processor architecture with simultaneous instruction issuing from multiple threads," *Proc. ISCA-15*, Los Alamitos, CA: IEEE Comput. Soc Press, 1988, p. 443.
13. A. Agarwal, B. H. Lim, D. Ktanz, and J. Kubiaticz, "APRIL: a processor architecture for multiprocessing," *Proc. ISCA-17*, Los Alamitos, CA: IEEE Comput. Soc. Press, 1990, p. 104.
14. W. J. Dally, S. W. Keckler, N. Carter, A. Chang, M. Fillo, and W. S. Lee, "M-Machine architecture v 1.0," *MIT Concurrent VLSI Architecture Memo 58*, MIT, Cambridge, MA, 1994.
15. S. J. Eggers, J. S. Emer, H. M. Levy, J. L. Lo, R. L. Stamm, and D. M. Tullsen, "Simultaneous multithreading: a platform for next-generation processors," *IEEE Micro. J.* 17, 12 (Sept/Oct. 1997).
16. G. Gao, K. Theobald, A. Marquez, and T. Sterling, "The HTMT program execution model," *Tech. Memo 09*, CAPSL, Univ. of Delaware, Newark, 1997.
17. M. Dorojevets, "The COOL ISA Handbook," *Tech. Report 04*, RSFQ System Group, SUNY, Stony Brook, NY, 1998.
18. P. Bunyk, and P. Litskevich, "Case Study in RSFQ Design: Fast Pipelined 32-bit Adder", *Reports 1998 Applied Supercond. Conf.*, Palm Desert, CA, Sept. 14-18, 1998, to be published in *IEEE Trans. Appl. Supercond.* (1999).
19. Y. Kameda, S. V. Polonsky, M. Maezawa, and T. Nanya, "Self-timed parallel adders based on DI RSFQ primitives," *Reports 1998 Applied Supercond. Conf.*, Palm Desert, CA, Sept. 14-18, 1998, to be published in *IEEE Trans. Appl. Supercond.* (1999).
20. S. Nagasawa, Y. Hashimoto, H. Numata, and S. Tahara, "A 380-ps, 9.5 mW Josephson 4-Kbit RAM operated at high bit yield", *IEEE Trans. Appl. Supercond.* 5, 2447 (1995).
21. D. Yu. Zinoviev, G. Sazaklis, S. Yorozu, and K. Likharev, "CNet: A superconductor network for petaflops computing," *Reports 1998 Applied Supercond. Conf.*, Palm Desert, CA, Sept. 14-18, 1998, to be published in *IEEE Trans. Appl. Supercond.* (1999).
22. D. Patterson and J. Hennessy, *Computer Architecture. A Quantitative Approach*, 2nd ed., San Francisco: Morgan Kaufmann, 1996.

Finite Frequency Shot Noise in Diffusive Wires

Yehuda Naveh

Department of Physics and Astronomy, SUNY at Stony Brook, Stony Brook, NY 11794-3800, U.S.A.

1. Introduction

Devices operating at low enough temperatures are not subject to thermal Johnson-Nyquist noise. On the other hand, shot noise, which is the current fluctuation due to finite voltage across the device, can be significant even at zero temperature. This type of noise is of particular importance at finite observation frequencies ω , where $1/f$ noise becomes small. Previously, shot noise was extensively studied for devices with tunneling components, such as the resonant tunneling diode¹⁻⁴ and single-electron tunneling structures.⁵⁻¹⁰ Much less attention was given in the device literature to shot noise originating in the conductors leading to the active region of the device, as well as in the conductors interconnecting these regions. Perhaps one reason for this apparent neglect of the interconnects as noise sources is the common belief that shot noise in solid state systems is appreciable only if the scale of the device is microscopic. In this work it is shown that this is not the case, and finite frequency shot noise may be appreciable in long, as well as in short, wires.

The origin of shot noise in ordinary diffusive conductors is quite different from the case of a vacuum diode. In the latter, noise is due to the random emission process at the cathode, whereas in diffusive conductors, random scattering events throughout the conductor contribute to the current fluctuations. These events can be strongly correlated due to electron-electron (e-e) interactions, manifested by short-range e-e scattering, by screening, and by the Pauli exclusion principle. Due to such correlations, it is known that the shot noise spectral density $S_I(\omega)$ may assume a value which is different from the classical Schottky value of $2eI$ (where I is the dc current), and which depends on the form of scattering in the conductor¹¹⁻¹⁵ and on frequency.¹⁶⁻¹⁸ However, as long as energy dissipation is small in the conductor, the shot noise value in all those cases remains of the order of $2eI$. When dissipation in the conductor is strong (i.e., when the conductor's length L is much larger than the electron-phonon relaxation length l_{ph}), zero frequency shot noise is strongly reduced, and tends to zero in the limit of $L \gg l_{ph}$.^{11,12,19} This decrease of the noise with L is particularly slow,^{12,20} which means that even low-frequency shot noise may be substantial in wires of length comparable to $1 \mu\text{m}$. Furthermore, it will be shown below that at typical device operating frequencies shot noise may not decrease with sample's length, but rather remain of the order of $2eI$ for any ratio of L and l_{ph} .

2. Model and theory

The results presented here are based on the "drift-diffusion-Langevin" theory developed in Refs. 16, 18 and valid at $\omega \ll 1/\tau$ and $\omega \ll eV/\hbar$, where τ is the elastic scattering time and V the voltage across the conductor. According to this theory, the noise spectral density as measured in the electrodes connecting the conductor can be expressed as

$$S_I(\omega) = \frac{2G}{L} \int_{-\frac{L}{2}}^{\frac{L}{2}} |K(x; \omega)|^2 C(x) dx \quad (1)$$

where G is the dc conductance, $C(x) = 2 \int f(x, E) [1 - f(x, E)] dE$ is the correlator of local fluctuations, and $f(x, E)$ is the local distribution function of electrons. For example, for an equilibrium Fermi-Dirac distribution with temperature T , $C(x) = 2T$, and the noise assumes the Johnson-Nyquist value. For typical non-equilibrium distributions, $C(x)$ becomes much larger than the equilibrium correlator, and at $T = 0$ the noise given by Eq. (1) is the shot noise.

The response function $K(x; \omega)$, which is solely responsible for the frequency dispersion of the noise, gives the current generated in the electrodes by a fluctuating unit current source at x . It is dependent upon the specific geometry of the conductor and its electrodynamic environment, but its integral over the sample length always equals unity. We will study here a simple and realistic geometry for which an analytical expression for $K(x; \omega)$ can be obtained. It involves a thin and long conductor close to a ground plane, and connected on both sides by two electrodes of resistance much smaller than that of the conductor. Those electrodes, in turn, may lead to the active components of the device. Thus the conductor in our problem is modeling a weak link in a network of wiring connecting the various devices. It is assumed that both thickness of the conductor (in the direction perpendicular to the ground plane) and its distance from the ground plane are much smaller than L . The most apparent realization of this geometry is a thin film close to a gate electrode. Then the response function is given by¹⁸

$$K(x; \omega) = \kappa \frac{L \cosh(\kappa x)}{2 \sinh(\kappa L/2)}. \quad (2)$$

Here $\kappa(\omega) = (-i\omega D')^{1/2}$, with $D' = D + GL/C_0$, where D is the diffusion coefficient and C_0 is the (dimensionless) linear capacitance between the conductor and the ground plane.

It is clear from Eq. (2) that the only current fluctuations that are of importance in inducing noise in the electrodes are those which are within a distance $\lambda_\omega = 1/|\kappa(\omega)|$ from the conductor-electrode interfaces. Therefore, at high enough frequencies, the measured noise is associated with the highly non-equilibrium distribution of electrons near the edges of the conductor, and not necessarily with the distribution in the bulk of the sample (in long samples, the latter can be very close to an equilibrium distribution, a fact which led to the common belief that the shot noise should vanish in long samples). This simple argument means that whenever λ_ω is smaller than some length scale l_s (which gives the spatial extent of

non-equilibrium electrons in the conductor), the shot noise value should remain significant even with increasing L .

3. Results

In order to give a quantitative description of the above effect, one should solve the Boltzmann equation for the non-equilibrium distribution f . To this end, we follow the prescription of Ref. 20, where f was calculated under the assumption of longitudinal acoustic phonon scattering. It was shown in that paper that the width of the layer in which the electron distribution is far from equilibrium is $l_s \approx (Ll_{ph})^{1/2}$, with l_{ph} the inelastic scattering length of an electron due to emission of a phonon of energy eV . Therefore, one should expect large shot noise if $\lambda_\omega \approx (D'/\omega)^{1/2} < (Ll_{ph})^{1/2}$, or $L > L_0(\omega)$ with

$$L_0(\omega) = \frac{D'}{l_{ph}\omega}. \quad (3)$$

In what follows we would be interested in relatively long samples and low frequencies. Therefore we assume here that the electron-electron scattering length is much shorter than L and λ_ω . Making use of numerical results for f (and thus for $C(x)$) in this situation, we can find the noise spectral density by combining Equations (1) and (2).

Results for the noise spectral density $S_N(\omega)$ are presented in Fig. 1 for a specific set of experimental parameters. The upper curves in the figure show the total noise. The lower curves show, on the same scale, the thermal noise. Since the latter is smaller by at least an order of magnitude than the former, the upper curves actually depict the shot noise.

The physical discussion presented above is fully supported by the results shown in Fig. 1. One sees that at each of the three frequencies depicted, the noise initially decreases with L up to $L \approx \lambda_\omega$, whereupon it increases, and reaches its mesoscopic value again at $L \approx L_0(\omega)$. The initial decrease of the noise with increasing L is due to the electrons being increasingly thermalized in the bulk of the sample, while the subsequent increase is due to the widening of the non-equilibrium surface layer as $(Ll_{ph})^{1/2}$, and therefore the increasing distance from equilibrium of the noise-inducing electrons within the layer of distance λ_ω from the interfaces. As expected, at strictly zero frequency the noise reduces monotonically to the thermal value at $L \rightarrow \infty$.

4. Discussion

The unusual result of shot noise increasing with increasing sample length is essentially due to a competition between two independent physical processes: screening and equilibration. The importance of screening in affecting shot noise was first discussed by Landauer in qualitative terms,²¹ and was later studied quantitatively in Refs. 16, 18. Its outcome effect is summarized by Eq. (2).

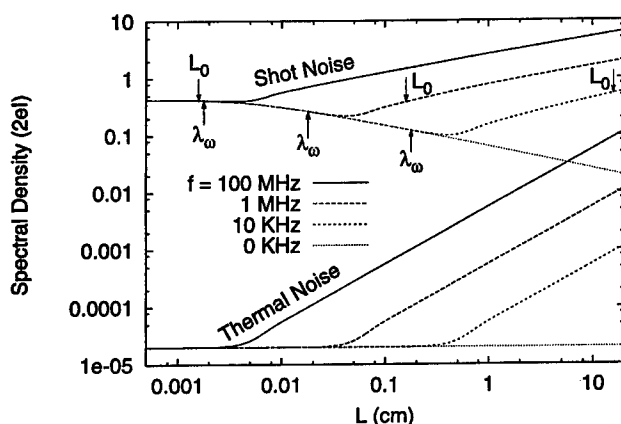


Figure 1. Noise spectral density vs. sample length L . Lower curves show the thermal noise and upper curves the full noise which is dominated by the shot noise. Material parameters are $l_{ph} = 10^{-3}$ cm and $D = 1000$ cm²/s. Temperature-to-voltage ratio is $\pi k_B T / eV = 10^{-5}$. Arrows indicate the positions of λ_ω and L_0 for each of the depicted frequencies (at zero frequency these lengths tend to infinity).

Equilibration, on the other hand is responsible for the surface layers of non-equilibrium electrons. The fact that the width l_s of these layers grows with L is readily understood:²⁰ since the electron-phonon relaxation time decreases strongly with the energy of the emitted phonon, at large L , when the electric field in the conductor is small, an electron entering the sample from the electrode will diffuse elastically for a long distance before emitting a phonon.

Parameters chosen in obtaining Fig. 1 correspond to a typical operating environment of future low-temperature microelectronic devices (e.g., single electron tunneling devices). At typical operating frequencies (higher than 1 MHz) one sees that shot noise remains of the order of $2eI$ whatever the length of the conductor. Moreover, even "zero frequency" experiments are invariably performed at an actual frequency of 10 kHz or higher, and should therefore reveal large shot noise when the sample is long. A different choice of the effective diffusion coefficient D' would leave the results unaltered if accompanied by a similar change of the frequency (in other words, the only dependence of S_I on ω and on D' is through ω/D'). While $D = 1000$ cm²/s is quite realistic, the electrostatic term in D' , $GL/C_0 \approx D(td/\Lambda_0^2)$, dominates if the thickness t of the conductor or its distance d from the ground plane are larger than the static screening length Λ_0 . Thus, the results depicted by Fig. 1 are of particular importance when the conductor is very thin, possibly a two-dimensional electron gas. With thicker conductors, the macroscopic shot noise would be large only at higher frequencies, or longer samples, as described by Eq. (3).

The shot noise discussed in this work is due to the voltage drop across the conductor in question (leads, interconnects, etc.). The resistance of such conductors, and thus the voltage, can become quite large as future cross-sections of the conductors diminish. However, that voltage can be significantly smaller than the actual voltage applied across the device. Moreover, for L_0 to be reasonably

small l_{ph} must be large. To maintain $l_{ph} = 10^{-3}$ cm, V cannot be larger than about 100 mV. It is then likely that in an actual situation T/eV would not be smaller than 10^{-3} . So, at large L and ω , thermal noise may be as large as the shot noise.

The response function of Eq. (2) and the other results presented here are not necessarily valid for geometries different from the one studied here. The question of whether any specific geometry exhibits shot noise when the conductor is long enough reduces to the question whether finite-frequency fluctuations in the bulk of the conductor are sufficiently screened as to not induce current in the electrodes. Theoretically, a detailed answer to this question may involve difficult solutions of the Poisson equation. However, in a charged Fermi system finite-frequency currents are known to be screened beyond some typical length scale λ_ω' which does not depend on L .²² On the other hand, the "hot-electron" length scale $l_s \approx (Ll_{ph})^{1/2}$ is independent of the specific geometry. Therefore, it can be argued that in sufficiently long samples of an arbitrary geometry l_s is larger than λ_ω' , so the only important sources of noise are from the non-equilibrium regions near the electrodes. Following the physical discussion below Eq. (2), this implies that the qualitative features of the results presented here may be of a general nature.

5. Conclusions

It was shown that at low temperatures shot noise originating in diffusive conductors leading current to the active part of the device may be significant. This effect is particularly strong when the conductors are thin, for example in a two-dimensional electron gas. Then, at typical operating frequencies (say, 1 MHz), shot noise is of the order of the Schottky value for any length of the conductor.

6. Acknowledgments

I am indebted to D. V. Averin and K. K. Likharev for many fruitful discussions. The work was supported in part by the DOE (DE-FG02-95ER14575).

References

1. Y. P. Li, A. Zaslavsky, D. C. Tsui, M. Santos, and M. Shayegan, "Noise characteristics of double-barrier resonant-tunneling structures below 10 kHz," *Phys. Rev. B* **41**, 8388 (1990).
2. J. H. Davies, P. Hyldgaard, S. Hershfield, and J. W. Wilkins, "Classical theory of shot noise in resonant tunneling," *Phys. Rev. B* **46**, 9620 (1992).
3. E. R. Brown, "Analytical model of shot noise in double-barrier resonant-tunneling diodes," *IEEE Trans. Electron Dev.* **39**, 2686 (1992).
4. G. Iannaccone, G. Lombardi, M. Macucci, and B. Pellegrini, "Enhanced shot noise in resonant tunneling: theory and experiment," *Phys. Rev. Lett.* **80**, 1054 (1998).

5. K. K. Likharev, "Single-electron transistors: electrostatic analogs of the dc SQUIDs," *IEEE Trans. Magn.* **23**, 1142 (1987).
6. D. V. Averin and K. K. Likharev, "Single electronics: a correlated transfer of single electrons and Cooper pairs in systems of small tunnel junctions," in: B. L. Altshuler, P. A. Lee, and R. A. Webb, eds., *Mesoscopic Phenomena in Solids*, Amsterdam: Elsevier, 1991.
7. A. N. Korotkov, D. V. Averin, K. K. Likharev, and S. A. Vasenko, "Single-electron transistors as ultrasensitive electrometers," in: H. Koch and H. Lubbig, eds., *Single Electron Tunneling and Mesoscopic Devices*, Berlin: Springer-Verlag, 1992.
8. H. Birk, M. J. M. de-Jong, and C. Schönenberger, "Shot-noise suppression in the single-electron tunneling regime," *Phys. Rev. Lett.* **75**, 1610 (1995).
9. A. N. Korotkov, "Langevin approach for the shot noise calculation in single-electron tunneling," *Europhys. Lett.* **43**, 343 (1998).
10. K. A. Matsuoka and K. K. Likharev, "Shot noise of single-electron tunneling in one-dimensional arrays," *Phys. Rev. B* **57**, 15613 (1998).
11. C. W. J. Beenakker and M. Büttiker, "Suppression of shot noise in metallic diffusive conductors," *Phys. Rev. B* **46**, 1889 (1992).
12. K. E. Nagaev, "On the shot noise in dirty metal contacts," *Phys. Lett. A* **169**, 103 (1992).
13. K. E. Nagaev, "Influence of electron-electron interaction on shot noise in diffusive contacts," *Phys. Rev. B* **32**, 4740 (1995).
14. V. I. Kozub and A. M. Rudin, "Shot noise in the mesoscopic 2D diffusive systems in the limit of strong electron-electron scattering," *Surf. Sci.* **361/362**, 722 (1996).
15. M. J. M. de-Jong and C. W. J. Beenakker, "Shot noise in mesoscopic systems," in L. L. Sohn, L. P. Kouwenhoven, and G. Schoen, eds., *Mesoscopic Electron Transport*, Dordrecht: Kluwer, 1997.
16. Y. Naveh, D. V. Averin, and K. K. Likharev, "Effect of screening on shot noise in diffusive mesoscopic conductors," *Phys. Rev. Lett.* **79**, 3482 (1997).
17. K. E. Nagaev, "Long-range Coulomb interaction and frequency dependence of shot noise in mesoscopic diffusive contacts," *Phys. Rev. B* **57**, 4628 (1998).
18. Y. Naveh, D. V. Averin, and K. K. Likharev, "Noise properties and ac conductance of diffusive mesoscopic conductors with screening," submitted to *Phys. Rev. B* (1998).
19. R. Landauer, "Solid-state shot noise," *Phys. Rev. B* **47**, 16427 (1993).
20. Y. Naveh, D. V. Averin, and K. K. Likharev, "Shot noise in diffusive conductors: a quantitative analysis of electron-phonon interaction effects," submitted to *Phys. Rev. B* (1998).
21. R. Landauer, "Qualitative view of quantum shot noise," *Ann. New York Acad. Sci.* **755**, 417 (1995); "Mesoscopic noise: common-sense view," *Physica B* **227**, 156 (1996).
22. D. Pines and P. Nozières, *The Theory of Quantum Liquids*, New York: Benjamin, 1966, Sec. 3.5.

Short-Channel AIM-SPICE Models for Amorphous Silicon and Polysilicon Thin Film Transistors

B. Iñiguez, L. Wang, Z. Xu, T. A. Fjeldly,* and M. S. Shur

Electrical, Computer and System Engineering Department, Rensselaer Polytechnic Institute, Troy, NY 12180, U.S.A. and

**Center for Technology at Kjeller, Norwegian University of Science and Technology, N-2007 Kjeller, Norway*

1. Introduction

Amorphous and polysilicon based giant area integrated circuits will lead to the development of novel ultra high-resolution large-area displays and possibly to the development of a new futuristic "computer-on-glass". They will also be used in a great variety of inexpensive and reliable consumer products, often driven by on-board solar cells, made from the same material. However, in order to realize this great potential, we have to learn how to design integrated circuit using these highly non-ideal devices that not only have a much larger parameter variation but also have a "history", changing their characteristics under light or after prolonged voltage application. This learning will require a new generation of efficient physics-based computer aided design (CAD) tools for device design and optimization, circuit design, parameter extraction, and quality assurance and control. In order to develop these tools, one has to marry the insight into the device physics with a robust mathematical description to insure convergence; to combine intuition, physics, and mathematics. Empirical and semi-empirical parameters in the device models will be unavoidable but they must be chosen in such a way that they are easily (and automatically) extractable and scale with geometry to ensure that the models have a predictive capability. This challenge springs from an emerging trend in device modeling, also applicable to deep submicron silicon and to proposed wide-bandgap semiconductor devices.

Most of the TFT models presented so far have been validated only for long-channel devices. However, as device sizes are shrunk, new effects arise that need to be accounted for in the models. In poly-Si TFTs, the floating-body effect, caused by electron-hole generation via impact ionization at high drain voltages, becomes important in the saturation and subthreshold regimes. Other small-geometry effects are the drain-induced barrier lowering (DIBL), which is stronger than in crystalline MOSFETs, and the mobility degradation at high gate voltages. In short-channel amorphous silicon (a-Si) TFTs, we have found that two new effects become relevant in saturation: self-heating and the floating-body effects. In these devices, self-heating raises the channel temperature as the channel current increases. This temperature increase results in an additional increase of current.

The floating-body effect originates from electron-hole generation by tunneling near the drain at high drain-source voltages, which results in a decrease of the threshold voltage and a strong increase of the drain current.

In this article we outline CAD models for short-channel poly-Si and a-Si TFTs that will be implemented in our circuit simulator, AIM-Spice. Our approach is based on the unified charge control modeling concept,¹ which allows us to obtain an accurate and continuous description through the transition between the various regimes of device operation. The present models, which include important short-channel effects, are major improvements over our previous long-channel TFT models.² They are physics-based and include a minimal parameter set, mostly related to the device structure and the fabrication process. The automatic scaling of model parameters allows us to accurately model a wide range of device geometries.

2. Amorphous TFTs

We have developed an analytical and physics-based SPICE model for the dc operation of short-channel a-Si:H TFTs. This model uses the same formulation as Ref. 2 for above and below threshold, but it includes two major short-channel effects: self-heating and the kink effect.

Self-heating raises the channel temperature as the channel current is increased. This leads to an increase of the number of free carriers, and therefore, to an additional increase of current.³ To deal with the self-heating effect, a thermal kinetic analysis is carried out and a physical model and an equivalent circuit are used to estimate the thermal resistance of the device. In deriving an analytical description for self-heating, a first order approximation and self-consistency are used to give an iteration-free model⁴ accurate for a temperature rise of up to 100 °C.

We determined the dependencies of the model parameters on channel temperature from experimental measurements. The parameters that are significantly affected by temperature are the threshold voltage V_T , the field-effect mobility μ_{FET} and the saturation voltage V_{sat} .⁵ Experimentally, we observed that as the temperature increases, V_T decreases while μ_{FET} and V_{sat} increase. Linear expansions of the expressions of these parameters in terms of the temperature rise ΔT allow us to establish an analytical formalism of the $I(V)$ model, including self-heating, to first order in ΔT . On the other hand, ΔT can be expressed as the ratio of the power dissipated in the channel and the device thermal impedance, R_{th} . Eliminating ΔT between these two considerations yields the following self-consistent expression for the drain current:⁶

$$I_D = I_{D0} / [1 - (I_{D0} V_{\text{DS}} / R_{\text{tot}} T_{\text{tot}})] \quad (1)$$

where I_{D0} is the drain current without self-heating, V_{DS} is the drain-source voltage, and T_{tot} is a thermal coefficient.

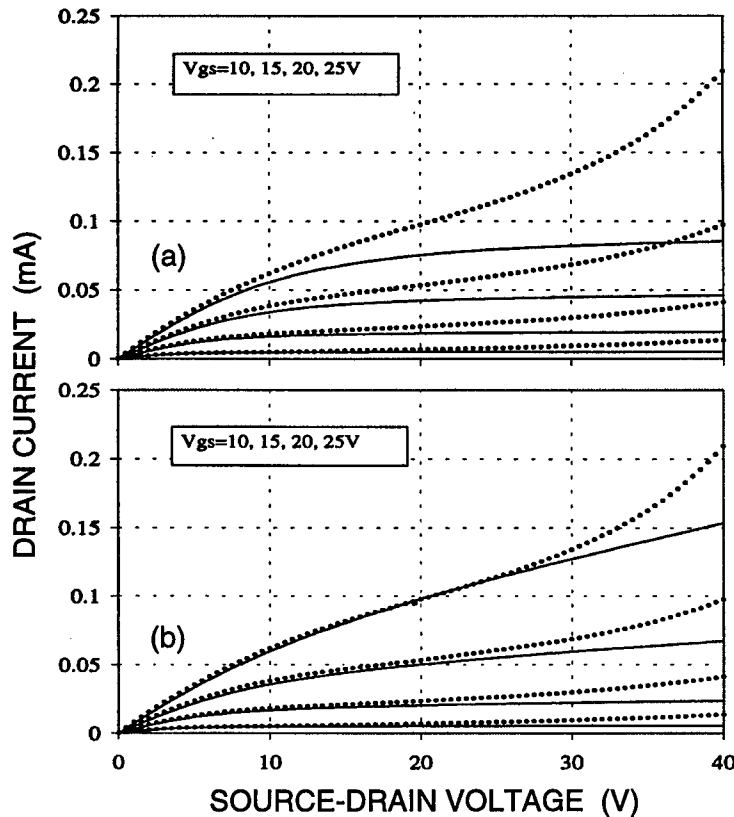


Figure 1. Modeled (lines) and measured (symbols) characteristics of an $L = 4 \mu\text{m}$ a-Si:H TFT. In (a) the model includes neither self-heating nor the kink effect; in (b) self-heating is included in the model but the kink effect is not.⁶

We extracted the device parameters using the automatic parameter extractor, AIM-Extract.⁷ Using the extracted device parameters and thermal resistances, we calculated the $I(V)$ characteristics for a $4 \mu\text{m}$ device using the basic device model with and without self-heating. Comparisons with the measured $I(V)$ characteristics for this device are shown in Fig. 1(a, b). By including self-heating, the validity of the model extends to drain biases of about 25 V. However, in Fig. 1(b), we observe a super-linear increase of the drain current at high drain biases, that cannot be explained by the self-heating theory. This "kink" effect is similar to that observed in poly-Si TFTs^{8,9} and in SOI MOSFETs,¹⁰ except that the hole generation mechanism in this case cannot be impact ionization, since the mean free path of the carriers in a-Si:H TFTs is very small, and therefore the carriers are unable to extract sufficient kinetic energy from the electric field. We believe that the kink effect in a-Si:H TFTs originates from tunneling from traps, which generates free electrons and holes if the lateral field is high enough. The generated

electrons and holes move to the drain and the source, respectively. The resulting hole current forward biases the potential barrier between the source and the adjacent depletion region of the a-Si film, causing an effective reduction in the threshold voltage and an increase in the channel current.

With the device biased in saturation, we assume that tunneling takes place only in the saturated part of the channel near the drain where the field is the highest. We can estimate the tunneling current using a Fowler-Nordheim expression.¹ The floating body effect, caused by holes generated in the tunneling process, appears to be the dominant contributor to the kink effect. To a first order approximation, the increase in the drain current is also proportional to the tunneling current and to the transconductance in saturation. Including the effects of self-heating and tunneling in the drain current, we obtain:⁶

$$I_D = I_{D0} / [1 - (I_{D0} V_{DS} / R_{tot} T_{tot}) + \Gamma V_{GT} (V_{DS} - V_{sat}) \exp[-V_K / (V_{DS} - V_{sat})]] \quad (2)$$

where V_{GT} is the gate voltage overdrive, and Γ and V_K are adjustable parameters.

A unified expression for the drain current, valid for all operating regimes, is obtained by replacing V_{sat} with an effective drain-source voltage and by replacing V_{GT} with an effective gate voltage overdrive V_{GTE} .¹

As shown in Fig. 2(a), the model given by Eq. (2) reproduces the experimental data quite well. For comparison, Fig. 2(b) shows modeled $I(V)$ characteristics for a hypothetical smaller $L = 1 \mu\text{m}$ a-Si:H TFT.

3. Polysilicon TFTs

We also developed a new model for poly-Si TFTs that is accurate for both long and short-channel devices. We included effects that were not considered in Refs. 8 and 9, but that are non-negligible even for long-channel devices. The model is based on the "effective medium" approximation that treats the polycrystalline material as a uniform medium with effective material properties — reasonable if the number of grain boundaries is sufficiently high. The model includes the dependencies of the parameters on the channel length to make it scalable.

The drain current model is based on the same compact formulation that has been used for other FETs,¹ using unified expressions for the channel charge and the mobility. The mobility model includes the degradation at high gate voltages caused by increased surface roughness scattering, and the effects of the increase in the threshold voltage at moderate inversion owing to a decrease of the potential barrier height of the grain boundaries.^{8,11} We also included the enhanced DIBL effect observed in poly-Si TFTs, resulting from the grain barrier height lowering caused by the lateral field, and from the depletion charge sharing between the gate-source and the gate-drain contacts.

The total drain current in poly-Si TFTs can be represented as the sum of three terms: a channel-current term that results from drift-diffusion transport, a kink effect term, and a leakage current.^{8,9} The kink-effect in this case is again related to

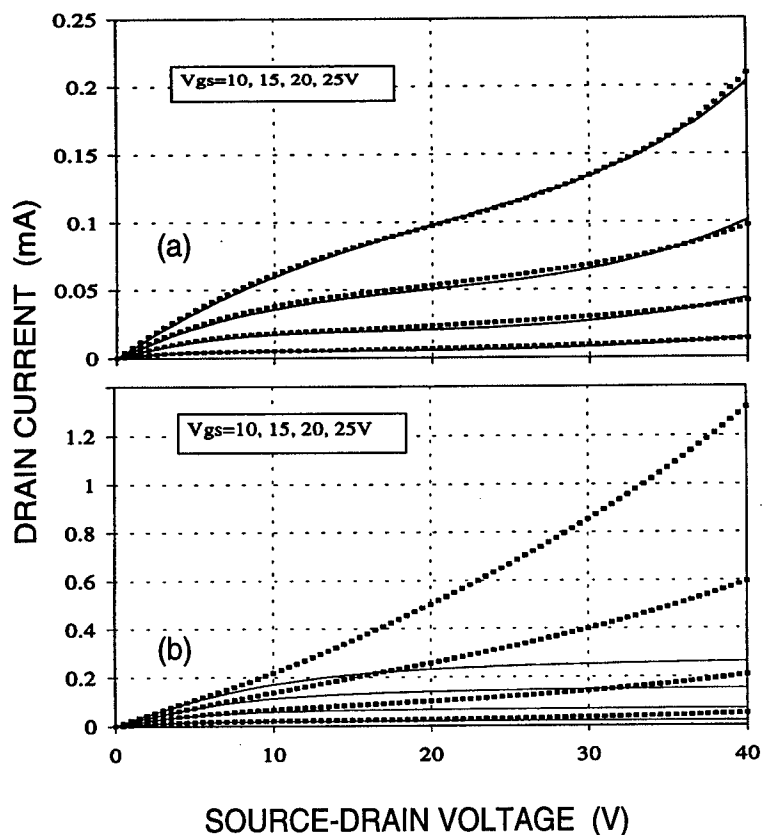


Figure 2. (a) Modeled (lines) and measured (symbols) characteristics of an $L = 4 \mu m$ a-Si:H TFT. The model includes both self-heating and the kink effect. (b) Modeled characteristics of an $L = 1 \mu m$ a-Si:H TFT with (symbols) and without (lines) self-heating and the kink effect.⁶

the floating-body effect caused by holes generated in the channel. However, in this case the holes are generated by impact ionization in the high-field region near the drain. As a result of the floating-body effect, the body-source voltage increases and the subthreshold ideality factor decreases with increasing drain-source voltage, similar to what has been observed in SOI MOSFETs.¹⁰

We extracted the first-order estimates of the device parameters by determining each parameter in the regime where it is relevant. Next, a global optimization was carried out to extend the validity of the parameter values to all operating regimes. A detailed explanation of the extraction procedure will be published elsewhere. Comparisons with the measured poly-Si TFT $I(V)$ characteristics of an $L = 2 \mu m$ device are shown in Fig. 3.

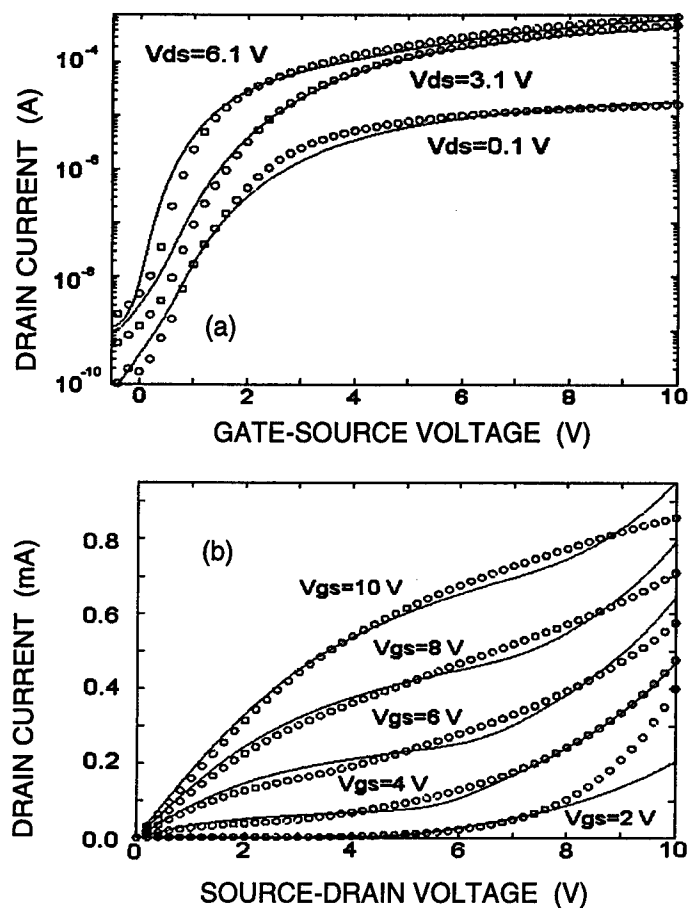


Figure 3. Measured (symbols) and modeled (lines) transistor characteristics for a device with dimensions $L = 2 \mu\text{m}$, $W = 10 \mu\text{m}$, and gate insulator thickness of 100 nm .¹²

4. Conclusions

We have discussed possible future applications of short-channel amorphous and polysilicon thin film transistors, and outlined novel physics-based models for such devices. The new models incorporate effects that are pronounced in short-channel devices, such as the kink effect in both a-Si and poly-Si TFTs and self-heating in a-Si TFTs. The model parameters include scaling functions that extend the validity of the model to a wide range of channel lengths. The poly-Si TFT model accurately reproduces the dc characteristics for channel lengths down to $2 \mu\text{m}$ in

all regimes of operation. The a-Si TFT model has been verified down to a 4 μm gate length. The new models agree well with experiments. They will be implemented in the circuit simulator AIM-Spice⁷ in the near future.

5. Acknowledgments

The authors would like to acknowledge the support from DARPA (Dr. G. Henderson, project monitor). We would like to thank Dr. Rene Lujan at dpiX Inc. for the sample preparation. We are also very grateful to Dr. Masakiyo Matsumura at Tokyo Institute of Technology, Dr. Richard Crandall at NREL, and Dr. Arthur Bergles at Rensselaer Polytechnic Institute for their help with information on amorphous silicon material properties. Finally, we would like to thank Mr. Andrew Dickens and Dr. Jianqiang Lu for help with the experimental set-ups.

References

1. T. A. Fjeldly, T. Ytterdal, and M. S. Shur, *Introduction to Device Modeling and Circuit Simulation*, New York: Wiley, 1998.
2. M. S. Shur, H. C. Slade, T. Ytterdal *et al.*, "Modeling and scaling of a-Si:H and poly-Si thin film transistors," *Mat. Res. Soc. Proc. Amorphous Microcrystal. Silicon Technol.*, Vol. 467, 1997, p. 831.
3. C. H. Kim and M. Matsumura, "Short channel amorphous-silicon thin-film transistors," *IEEE Trans. Electron Dev.* **43**, 2172 (1996).
4. Y. Cheng and T. A. Fjeldly, "Unified physical *I-V* model including self-heating effect for fully depleted SOI/MOSFETs," *IEEE Trans. Electron Dev.* **42**, 1291 (1996).
5. H. C. Slade, *Device and Material Characterization and Analytic Modeling of Amorphous Silicon Thin Film Transistors*, Ph.D. Dissertation, Univ. of Virginia, 1997.
6. L. Wang, T. A. Fjeldly, B. Iñiguez, H. C. Slade and M. S. Shur, "Self-heating and kink effects in a-Si:H thin-film transistors," submitted to *IEEE Trans. Electron Dev.* (1998).
7. See the AIM-Spice WWW homepage, <http://www.aimspice.com>.
8. A. Owusu, M. Jacunski, M. S. Shur and T. Ytterdal, "SPICE model for the kink effect in polysilicon TFTs," *Abstracts 1996 Fall Meeting Electrochem. Soc.*, Vol. 96-2, San Antonio, TX, October 6-11, 1996.
9. M. Jacunski, *Characterization and Modeling of Short Channel Polysilicon Thin Film Transistors*, Ph.D. Dissertation, Univ. of Virginia, 1996.
10. D. Suh and J. G. Fossum, "A physical charge-based model for non-fully depleted SOI MOSFETs and its use in assessing floating-body effects in SOI CMOS circuits," *IEEE Trans. Electron Dev.* **42**, 728 (1995).

11. M. Jacunski, M. S. Shur, A. A. Owusu, T. Ytterdal, and M. Hack, "SPICE models for n and p channel polysilicon thin film transistors in all regimes of operation," in: *Proc. AMLCDs '95 Workshop*, September 1995, p. 134.
12. B. Iñiguez, Z. Xu, T. A. Fjeldly, and M. S. Shur, "Unified model for short-channel poly-Si TFTs," submitted to *Solid State Electronics* (1998).

3 Alternative Paths to Nanoelectronics: Self-organization, Molecular Engineering

This chapter covers one of the most exciting alternative paths in the development of microelectronics, self-organization and molecular electronics. Over its life span, microelectronics has always evolved along the miniaturization path — from the larger to the smaller. However, further progress towards nanoelectronics along this path appears to promise exponentially increasing costs and diminishing returns. In fact, some of the analyses presented in Chapter 1 project the possibility of zero or even negative return as the increasing bit cost curve crosses over the decreasing bit price curve. Looking for salvation in bold moves to larger wafers and greater complexity, we must admit that, however effective these moves may be, they will require even greater investments on top of the already barely affordable fab costs. Is this really the only way to continue? Do we have to approach the realm of nanoelectronics from that of microelectronics? What if that realm can, in fact, be reached from the opposite direction — from atoms and molecules — at a much-reduced cost?

A rather unexpected analogy devised by one of the Workshop's more provocative participants compared the present microelectronics to Stone Age construction work. At the outset, our ancestors built their shelters by clearing caves and digging into the ground. Yet, with time, we learned to build using "molecular" units, like bricks and mortar. Similar new ways of "building up" are highlighted in the following chapter. Some of the promising new directions are self-assembly of quantum dots, non-lithographic fabrication of nanowire and nanodot arrays based on self-organization of pores in anodization, and molecular self-assembly.

This chapter opens with an article by Nikolai Ledentsov, describing the very exciting development of quantum dot lasers fabricated using self-organization phenomena at crystal surfaces during epitaxial growth. While presenting his paper at Embiez, Ledentsov argued eloquently that the *commercial advent* of quantum dot lasers is imminent. Horst Stormer, a skeptic about the immediate prospects of this novel technology, challenged him on that point and a \$1,000 bet was wagered and accepted. One of these distinguished gentlemen will have to pay up on or before the year 2003. It is worth noting that Stormer has already begun building a war chest by collecting scientific prizes of significant cash value.

Quantum Dot Lasers: Experimental Results and Future Trends

N. N. Ledentsov

*A. F. Ioffe Physical-Technical Institute, 194021, St. Petersburg, Russia and
Institut für Festkörperphysik, TU Berlin, D-10623 Berlin, Germany*

1. Introduction

A semiconductor crystal with a size of only several nanometers, or quantum dot (QD), keeps the basic properties of the atom yet provides a geometrical size that allows experiments of atomic physics to be held on a relatively macroscopic object.

In device applications, QDs allow the semiconductor engineer to fulfill the challenging task of fabricating devices that operate with the same good performance at both high and low temperatures. When the first papers by Arakawa and Sakaki¹ and by Asada *et al.*² on the possibility of using QDs as the active media of semiconductor lasers with greatly improved and temperature-insensitive parameters appeared in the early 1980s, many scientists and engineers started searching for ways of fabricating quantum dots (QDs) and studying their properties. However, more than a decade passed until the first QD lasers were fabricated in 1993³ and proven to demonstrate the predicted properties.⁴

It is well known that a single atom has discrete energy levels separated by forbidden energy gaps. When an atom is excited, the electron goes to the higher energy level. When it relaxes back to the ground state, a photon with a strictly defined energy is emitted.

Unlike the case of a dilute gas of atoms, the atoms in crystals are strongly bound to each other. The interactions between the closely spaced atoms in crystals make broadening of the electronic spectrum unavoidable. The absorption band becomes rather broad, on the order of several electron volts, in marked difference with the sharp line absorption spectra of single atoms.

The wide bands of allowed states in the crystal provide a lot of possibilities for scattering of electrons and holes. At high temperatures, lattice vibrations (phonons) can easily stimulate transitions of charge carriers in the energy range defined by the lattice temperature and/or scatter their direction. Generally, a "tail" of the carrier distribution near the bottom of the conduction band and the top of the valence band increases remarkably with temperature. For the same average concentration of injected carriers, the broadening of the energy spectra with increased temperature decreases the maximum gain and degrades the laser performance, among other disadvantages.

The situation changes remarkably if the motion of the charged carrier in the crystal is limited to a very small volume, e.g. in a three-dimensional rectangular

box. Localization of the carriers can be provided by the walls of the box or by its interfaces with a surrounding (matrix) material. In the latter case it is important that the matrix material provide a larger bandgap than the QD material and that the potential wells be attractive both for electrons and holes. If the size of the box is small, the electron energy spectrum is quantized, similar to the case of energy quantization in the attracting Coulomb potential of the atomic nuclei. In the simplified case of infinite barriers at the QD-matrix interface, the size quantization energy is described by:

$$E_{xyz} = \frac{\hbar^2 \pi^2}{2m_e^*} \left\{ \left(\frac{n_x}{L_x} \right)^2 + \left(\frac{n_y}{L_y} \right)^2 + \left(\frac{n_z}{L_z} \right)^2 \right\} \quad (1)$$

where m_e^* is the electron effective mass, E_{xyz} is the size quantization energy due to electron localization in a box with dimensions L_x , L_y , and L_z , respectively, and $n_{x,y,z} = 1, 2, 3, \dots$.

Electrons in crystals usually have rather small effective masses, so that a relatively large box size of about 10 nm results in a large energy separation between electron sublevels (about 100 meV for a GaAs QD). The latter value significantly exceeds the thermal energy at room temperature (26 meV) and thus the population of excited states can be avoided. This idea served as the basis of the QD laser concept in Ref. 1. As we will see later, a lot of other unexpected effects provide QD lasers with additional significant advantages.

One should note, however, that there was a lot of skepticism concerning the possible application of QDs to real devices. Traditional methods of QD fabrication based on patterning of structures with ultra-thin layers suffered either from insufficient lateral resolution or introduced heavy damage in the material upon processing.⁵ There were also predictions that, even if the ideal QD could be fabricated, it could hardly be applied to real devices, as ultra-long relaxation times between electron sublevels were expected.⁶

2. Self-organized growth of quantum dots

A solution for fabricating QDs with the required properties came from an effect that was traditionally considered as undesirable by crystal growers. It was found that if one deposits a layer of a material having a different lattice constant from that of the substrate, the growth of strained material first proceeds in a planar mode and a so-called "wetting" layer is formed. However, at some critical thickness this planar growth stops and three-dimensional nanoscale islands on top of the thin wetting layer are formed, as was demonstrated for the case of the growth of indium arsenide on gallium arsenide in 1985 by the group at CNET, France.⁷ When these islands are covered with GaAs, a GaAs pie with InAs raisins is formed. As the InAs has a much smaller bandgap energy than the GaAs, an array of InAs QDs is formed.

Initially this technique did not get much attention as the possibility of producing dislocation-free QDs of uniform size and shape was not evident. Just a

couple of years ago the idea of fabricating high-performance quantum dot lasers using the islanding effect was met with skepticism even by leading experts in this field.⁸ However, recent developments point to a much more optimistic outlook.

The driving force for the formation of three-dimensional islands is related to the elastic strain relaxation. The material on top of the pyramid can relax, expanding in vacuum to lower its stored elastic energy. For a 45° facet angle of the pyramid, 60% of the elastic energy accumulated in the biaxially compressed flat layer is relaxed.^{9,10} On the other hand, the formation of pyramids results in an increase in the total surface area. If the formation of islands results in an increase in the surface energy of the system, the initially formed islands will undergo ripening, as the system will try to reduce the total surface area covered by QDs.

An opportunity for the fabrication of uniform in size and in shape QDs stable with respect to ripening appears only if the total surface energy of the island is smaller than the surface energy of the corresponding area of the wetting layer occupied by it. If one takes into account that the major surface properties, e.g. surface reconstruction and surface stress, are strongly affected by the strain state of the crystal, one can conclude that they can differ significantly for the strained wetting layer and for the facet of the relaxed pyramid. Numerical estimations of the strain-induced renormalization of the surface energy made by Shchukin *et al*¹⁰ suggest that the formation of "equilibrium" equisized and equishaped islands that do not undergo ripening is, fortunately, probable.

The optimization of growth parameters to realize equilibrium arrays is a difficult task to be solved for each materials system.¹¹ If islands uniform in size and in shape are formed, one speaks about "self-organized quantum dots," as this system represents a clear example of the *spontaneous formation of macroscopic order from initially random distributions*. In the case of dense arrays of QDs, their interaction via the strained substrate makes their *lateral ordering* favorable.¹⁰ Growth on patterned surfaces can also give ordering of QDs.¹¹ For multi-stack QD deposition vertically-correlated growth of QDs has been demonstrated^{7,11,13} and, thus, *quasicrystals composed of quantum dots* either in two or in all three dimensions can be fabricated. For islands having a two-dimensional shape either correlated or anticorrelated growth is possible depending on the relative thickness of the spacer layer.¹⁴

Several promising ways to fabricate QDs using self-organization phenomena at crystal surfaces and in the bulk have been demonstrated (see Ref. 15 and references therein):

- spontaneous quasiperiodic faceting of crystal surfaces and heteroepitaxial growth on faceted surfaces;
- spontaneous phase separation in semiconductor alloys during growth or slow cooling;
- spontaneous alloy decomposition upon high-temperature annealing;

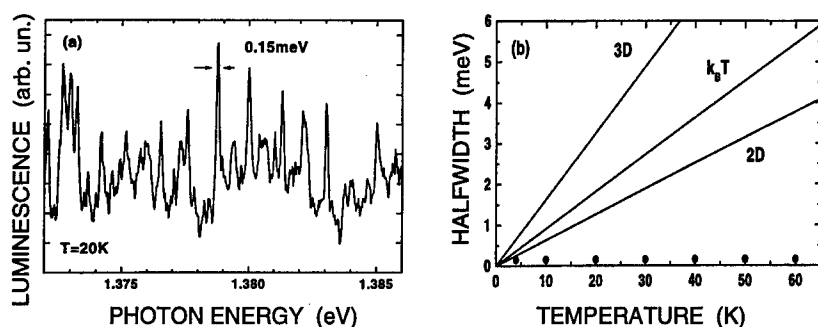


Figure 1. Ultra-sharp luminescence lines from single QDs (a) which do not show broadening with temperature increase (b).

- formation of two-dimensional islands during submonolayer heteroepitaxial deposition.

To date, most results have been obtained for QDs formed by three- and two-dimensional islanding during growth in lattice mismatched systems (see Ref. 16 and references therein).

3. Demonstration of an electronic quantum dot

An important breakthrough in the understanding of the properties of semiconductor QDs occurred when it was demonstrated that ultra-narrow luminescence lines from single InAs QDs^{17,18} exhibit no broadening with temperature,¹⁹ a very unusual phenomenon for any bandgap emission in semiconductors, but in full agreement with theoretical predictions for electronic QDs (see Fig. 1).

Much progress has been achieved in recent years, more than in the decades of previous research. We have learned about QDs in different materials systems, about the electronic spectrum in QDs, radiative recombination, and relaxation processes. Numerous teams are contributing to the development of this subject, among them CNET in France, UCSB, USC, Stanford University, Lund University, teams from Sheffield and Nottingham Universities, from Max-Planck Institute in Stuttgart and many other groups and institutions.

4. Expected results: edge-emitting and vertical cavity QD lasers

Evident progress in the use of QDs has been achieved in the area of semiconductor lasers. Two basic device geometries have been applied. In one case, the light propagates along the plane containing QDs and the resonator comprises a conventional Fabry-Perot cavity with natural cleaved mirrors (see Fig. 2, on the

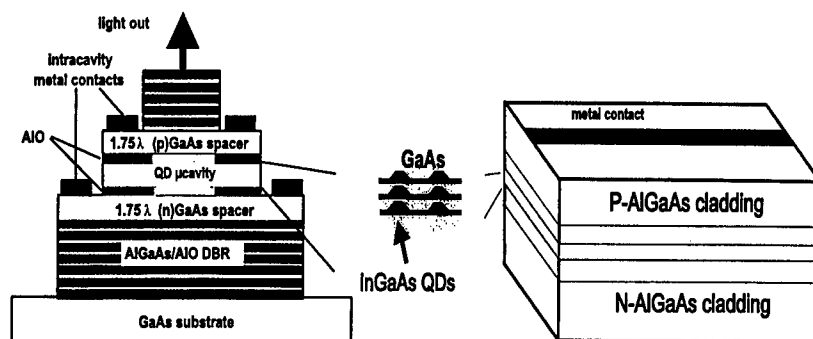


Figure 2. QDs are used as active media of semiconductor heterostructure lasers in edge-emitting (right) and vertical cavity (left) geometry.

right). In the other case, the light is emitted perpendicular to this plane (see Fig. 2, on the left), while the cavity is confined in the vertical direction by multilayer stacks of layers forming distributed Bragg reflectors (DBRs). The first approach allows the fabrication of high power lasers utilizing the advantage of ultra-low threshold current density due to QDs, and possibly the prevention of dislocation growth and the suppression of laser mirror overheating by nonradiative surface recombination due to localization of carriers in QDs.^{20,21} In the second approach lasers with ultra-low *total* currents can be fabricated and, even more exciting, lasers based on a *single* QD potentially can be realized.

Some of the important events in the QD laser field can be briefly listed here. The first photopumped QD laser was realized by Ledentsov *et al* in 1993.³ The first QD injection laser and lasing via the QD ground state exhibiting a temperature-insensitive threshold current came half a year later.⁴

Room temperature (RT) operation via quantum dots was demonstrated subsequently,²²⁻²⁶ as well as ultrahigh material and differential gain in QD lasers, which allowed for RT lasing at 60 A/cm² current density.²⁷ Continuous wave RT high power operation of a QD laser with 1.5 W output power was realized — see Fig. 3, with the device showing comparable performance to state-of-the-art quantum well lasers.²⁸

Other recent achievements include low-threshold InAs QD lasers on InP substrates emitting at 1.84–1.9 μm,²⁹ QD lasers emitting in the visible range,^{20,30} and high-performance QD vertical cavity surface emitting lasers (VCSELs),^{31,32} illustrated in Fig. 4. At the same time, significant progress in the theoretical understanding of QD lasers with realistic parameters has been achieved.^{33,34}

Generally, the basic parameters of edge emitting and vertical cavity QD lasers have approached the level of the devices based on QWs, currently dominating the market, *while the QD laser story is only starting.*

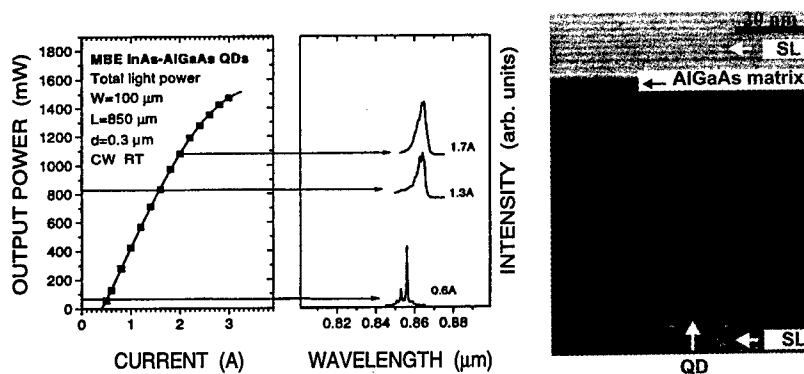


Figure 3. High-power operation of edge-emitting QD laser (left). Transmission electron micrograph of the active region of the high power QD laser (right).

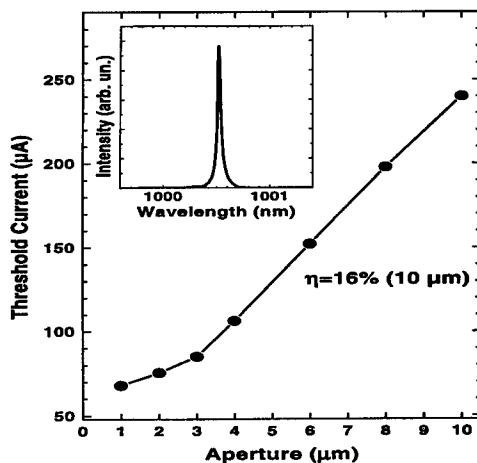


Figure 4. Threshold current of a QD VCSEL. The emission spectrum at $J = 1.3J_{th}$ is shown in the inset.

5. Unexpected results

• Far-infrared emission in quantum dot lasers.

New and unexpected applications in the laser field arise due to the discrete energy spectrum in QDs. In ultra-thin layers, or quantum wells, there exists a continuum of states at any energy above the subband energy, as the in-plane motion of charge carriers is not limited. If the carrier is excited to the second subband, it relaxes to the first subband via emission of discrete quanta of energy — optical (LO) phonons. Due to the continuum nature of electron states in a QW,

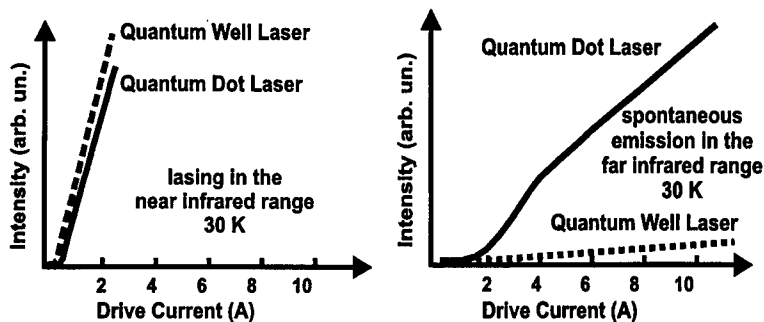


Figure 5. QD laser emits at $1\ \mu\text{m}$. Simultaneously, emission near $20\ \mu\text{m}$ appears.

there always exist states in the first subband to which electrons can scatter. That's why the electron lifetime in the second subband appears to be short (about 1 ps). The relaxation rate is reduced only near the subband edge, where no LO-phonon can be emitted, and the total electron cooling time usually takes about 20-100 ps. On the other hand, in the QD case the population time of the ground state and the depopulation time of the excited state coincide. However, due to the lack of the matching energies for electrons emitting LO phonons, the relaxation time to the ground sublevel takes typically 20-40 ps. The electron needs to emit a combination of different phonons to match the energy difference.¹⁷ Thus, the total relaxation time is comparable for QWs and QDs, while the excited state depopulation time is longer in QDs. This *increases the relative importance of the competing relaxation mechanism* via emission of *far-infrared (FIR) photons*.

In Fig. 5 we show light-current characteristics of a QD laser in the optical (upper curve) and in the FIR range by Vorob'ev *et al.*³⁵ The FIR emission from a conventional QW laser having similar geometry and threshold current is also shown by the dashed line. One can see that, in agreement with the difference in electron relaxation times in QWs and in QDs, the intensity of the FIR spontaneous emission is about one order of magnitude *higher* in the QD case. Moreover, the ground states of quantum dots depleted by lasing can't be filled by lateral diffusion of nonequilibrium carriers, the process that reduces a hole burning effect in quantum wells. The FIR emission in QDs has a threshold character as it requires fast depopulation of the QD ground state via stimulated emission in the optical range. For the QW case, empty states in the ground subband always exist (see Fig. 5), and weak FIR emission shows a non-threshold behavior. Much higher intensity of the FIR emission in the QD case, hopefully, will make it possible to create a new generation of FIR lasers.

- *Light emitting devices based on III-V quantum dots in silicon*

Another field where application of QDs can play an extraordinary role is related to silicon. Silicon, although being the major material for modern microelectronics, provides very low probability for injected electrons and holes to recombine radiatively and can hardly be used in lasers and light-emitting devices.

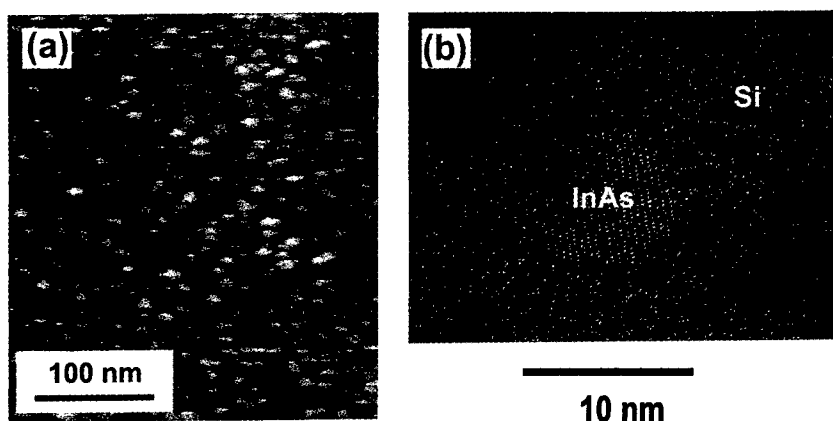


Figure 6. STM image of free-standing InAs QDs on Si (100) surface (a) and high-resolution transmission electron microscopy image of the Si-covered coherent InAs QD in a Si matrix (b).

If it is possible to insert narrow-gap QDs (e.g. made from InAs), which have high probability of radiative recombination, in such a way that electrons and holes will be trapped in these QDs, a silicon-based device with extremely efficient radiative recombination can be created. Recently, the possibility of depositing such InAs quantum dots on a Si surface was demonstrated.³⁶ A scanning tunneling microscope image of free-standing InAs QDs on a silicon (100) surface and a high-resolution transmission electron microscope image of silicon-overgrown InAs QDs are presented in Fig. 6a and b, respectively. Coherent InAs in a silicon matrix demonstrate broad PL emission peaking at 1.3 μm at 77 K and at 1.6 μm at RT, as shown in Fig. 7. We note also that introduction of InAs layers in a Si matrix, acting as barriers for electrons in this case, seems to be possible as well.

- *Extension of the spectral range of GaAs-based devices to 1.8 μm .*

The important point for laser technology is to have the possibility of covering the strategically important spectral ranges of 1.3 and 1.55 μm using GaAs substrates. This point is particularly important for VCSELs where high quality monolithic AlAs-GaAs Bragg reflectors and a well developed oxide technology are available only on GaAs substrates. QDs emitting at 1.3 μm at room temperature have been realized by several teams (see e.g. Ref. 15). Recently we discovered that complexes of InAs quantum dots are formed at low substrate temperatures.³⁷ These structures emit up to 1.8 μm at RT. Modifications of the growth mode provide the possibility of adjusting the luminescence to the 1.55 μm wavelength range.

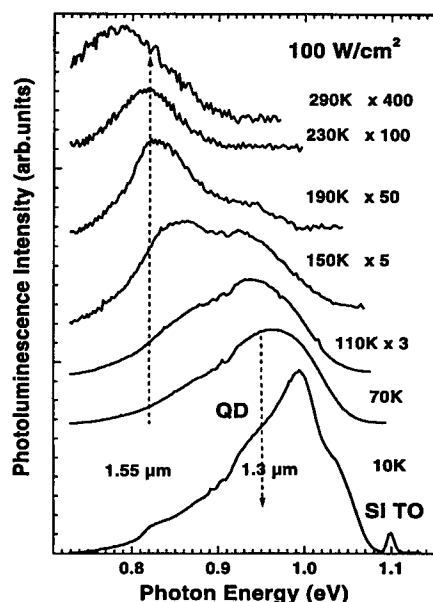


Figure 7. PL spectra of InAs QDs in a Si matrix.

- *Resonant waveguides*

Both flat (two-dimensional) and pyramid-like (three-dimensional) islands are currently used in devices. Despite the localization energy of excitons in 2D islands being smaller than in the case of 3D islands, the 2D islands demonstrate very high areal density and narrow luminescence and optical reflectance peaks^{6,18} allowing the realization of ultrahigh absorption coefficients. The experimentally measured absorption coefficient for structures with stacked CdSe QDs in a ZnSe matrix in the direction perpendicular to the planes with nanoislands approaches 10^5 cm^{-1} .³⁸ High absorption coefficients and the lack of exciton screening in dense arrays of QDs must result in ultrahigh quantum dot exciton (or, even higher, biexciton) gain values under the generation of nonequilibrium carriers.^{30,31} The latter is possible in QDs as the continuous in-plane motion of excitons is not possible and the related k -selection rule is broken³⁹ making the zero-phonon excitonic gain mechanism possible.

Resonant waveguides are based on the effect of resonant enhancement of the refractive index along the contour of the absorption (or gain) curve. To have a significant impact on the waveguiding properties, the absorption peak must be strong enough ($\Delta n \sim 0.5$ for $\alpha \sim 10^5 \text{ cm}^{-1}$).⁴⁰ This effect was experimentally measured in Ref. 41. For resonant waveguiding it is not necessary to have external cladding of the active region with QDs by layers with significantly lower refractive indices. Practically, it means that lasers can be created in materials having no suitable lattice-matched heterostructure with lower refractive index.

- *Self-adjusted cavities*

In VCSELs the strong resonant modulation of the refractive index serves for self-adjustment of the cavity mode and lasing spectrum. As the material gain of a single QD reaches ultrahigh values, even single quantum dot lasing may become possible.⁴⁰

- *Vertical cavity lasers without Bragg reflectors and cavity*

In edge emitting lasers a reflective coating is usually deposited on one facet to increase the output for the other facet, where an antireflection coating can be used. In contrast, highly reflective Bragg mirrors on both sides of the cavity are necessary for QW VCSELs, because of the relatively small maximum gain in these structures.² However, if the maximum gain can be made high enough, one Bragg mirror can be left out and the other need not necessarily be highly reflective. Thus for a gain value exceeding 10^5 cm^{-1} and an active layer thickness of 200 nm, a facet (or mirror) reflectivity of the order of 30% is enough to achieve vertical lasing. Due to the low finesse of the cavity and the self-adjustment effect no strict necessity of fitting the cavity mode and the gain spectrum exists, as was demonstrated for 20 stacked CdSe submonolayer QD insertions in a ZnMgSSe matrix⁴² grown on a GaAs substrate.

- *Quantum dot composites*

The gain of the array of QDs is not given by a simple sum of single QD gains. The interaction of electromagnetic fields of anisotropic QDs or anisotropic QD lattices makes the splitting of the TE and TM modes for the same QD exciton transition unavoidable. This effect can result in splitting as large as several tens of meV, as was predicted theoretically⁴³ and proven experimentally.⁴² The maximum gain of a QD array is also a strong function of the relative arrangement of QDs.⁴³

- *Quantum dots are everywhere*

Most modern industrial QW lasers are based on thin layers of alloys used as active regions. It has become clear now that these layers in most cases exhibit quasiperiodic nanoscale compositional modulations, creating in many cases dense arrays of quantum wire- or QD-like structures.¹⁵ For the same *average* alloy composition, the luminescence peak can be tuned by several hundreds meV by changing the growth conditions.⁴⁴ Careful evaluation of the impact of such effects on the lasing characteristics of modern lasers is necessary to clear up the role of the QDs in this case.

6. Conclusions

The appearance of QDs changed all the basic commandments of the double heterostructure (DHS) laser.⁴⁵ Let us compare the basic requirements for DHS (QW DHS) and QD lasers:

DHS and QW DHS lasers

- lattice matching
- material gain
- exciton screening
- homogeneous broadening at RT
- cladding with low n layers
- VCSEL: Bragg reflectors and cavity
- lasing in optical and near IR range
- one family (III-V on III-V, ...)
- limited wavelength range on GaAs

QD lasers

- undesirable
- orders of magnitude higher
- is not important
- is small
- is not necessary
- are not necessary
- and simultaneous FIR emission
- is not necessary (InAs/Si QDs)
- is extended to 1.8 μm

The QD laser seems to be a completely new device with properties that can remarkably expand our possibilities in many applications, rather than simply a laser with some parameter improved with respect to the DHS or DHS QW laser.^{1,2}

The much greater flexibility of QD lasers makes it very probable that these devices in the future will replace their more conventional prototypes. This idea still raises, however, a lot of skepticism.^{8,46} We note, however, that the idea of the DHS laser⁴⁵ was also considered as a "schreibtisch" patent at one point. Later, the advantages of QW lasers were doubted and it took seven years after the first demonstration until the competitive QW device was created. The possibility of using strained QWs in lasers was also doubted. Of course the possibility of fast progress in QD lasers depends on the general QD research area and also on industrial interest and support. One should consider competition with other directions, e.g. with polymer LEDs and lasers.

7. Acknowledgements

Original results discussed here were obtained in cooperation with my colleagues at A. F. Ioffe Physical-Technical Institute, Technische Universität Berlin, Technical University of St. Petersburg, Max-Planck-Institut für Mikrostrukturphysik (Halle), and the Institute for Analytical Instrumentation in St. Petersburg. Financial support from the Russian Foundation of Basic Research, Volkswagen Foundation, INTAS, and Deutsche Forschungsgemeinschaft and fruitful discussions with Zh. I. Alferov, D. Bimberg, G. E. Cirlin, L. Eaves, D. Gerthsen, M. Grundmann, M. Henini, J. Lott, M. V. Maximov, J. Merz, V. A. Shchukin, M. Skolnick, V. M. Ustinov, L. E. Vorob'ev and P. Werner are acknowledged. The author is personally grateful to the Alexander von Humboldt Foundation.

References

1. Y. Arakawa and H. Sakaki, "Multidimensional quantum well lasers and temperature dependence of its threshold current," *Appl. Phys. Lett.* **40**, 939 (1982).
2. M. Asada, M. Miyamoto, and Y. Suematsu, "Gain and the threshold of three dimensional quantum dot lasers," *IEEE J. Quantum Electron.* **QE-22**, 1915 (1986).
3. N. N. Ledentsov, V. M. Ustinov, A. Egorov *et al.*, "Optical properties of heterostructures with InGaAs-GaAs quantum clusters," *Semiconductors* **28**, 832 (1994).
4. N. Kirstaedter, N. N. Ledentsov, M. Grundmann *et al.*, "Low threshold, large T_0 injection laser emission from (InGa)As quantum dots," *Electron. Lett.* **30**, 1416 (1994).
5. S. P. Beamount, "Quantum wires and dots: the challenge to fabrication technology," in: A. R. Peaker and H. G. Grimmeiss, eds., *Low-Dimensional Structures in Semiconductors*, New York: Plenum Press, 1991, p. 109.
6. H. Benisty, C.M. Sotomayor Torres, and C. Weisbuch, "Intrinsic mechanism for the poor luminescence properties of quantum-box systems," *Phys. Rev. B* **44**, 10945 (1991);
T. Inoshita and H. Sakaki, "Electron relaxation in a quantum dot: significance of multiphonon processes," *Phys. Rev. B* **46**, 7260 (1992).
7. L. Goldstein, F. Glas, J. Y. Marzin, M. N. Charasse, and G. LeRoux, "Growth by molecular beam epitaxy and characterization of InAs/GaAs strained-layer superlattices," *Appl. Phys. Lett.* **47**, 1099 (1985).
8. See e.g. B.G. Levi "Researchers vie to achieve a quantum dot laser," *Physics Today*, May 1996, p. 22.
9. M. Grundmann, O. Stier and D. Bimberg, "InAs/GaAs pyramidal quantum dots: strain distribution, optical phonons, and electronic structure," *Phys. Rev. B* **51**, 11969 (1995).
10. V. A. Shchukin, N. N. Ledentsov, P. Kop'ev, and D. Bimberg, "Spontaneous ordering of arrays of coherent strained islands," *Phys. Rev. Lett.* **75**, 2968 (1995).
11. N. N. Ledentsov, M. Grundmann, N. Kirstaedter *et al.*, "Ordered arrays of quantum dots: formation, electronic spectra, relaxation phenomena, lasing," *Solid State Electron.* **40**, 785 (1996).
12. D. S. L. Mui, D. Leonard, L. A. Coldren, and P. M. Petroff, "Surface migration induced self-aligned InAs islands grown by molecular beam epitaxy," *Appl. Phys. Lett.* **66**, 1620 (1995).
13. Q. Xie, A. Madhukar, P. Chen, and N. Kobayashi, "Vertically self-organized InAs quantum box islands on GaAs (100)," *Phys. Rev. Lett.* **75**, 2542 (1995).
14. V. A. Shchukin, D. Bimberg, V. G. Malyshkin and N. N. Ledentsov, "Vertical correlations and anticorrelations in multisheet arrays of two-dimensional islands," *Phys. Rev B* **57**, 12262 (1998).
15. N. N. Ledentsov, "Self-organized quantum wires and dots: new opportunities for device applications," *Prog. Crystal Growth Charact.* **35**, 289 (1997).

16. R. Nötzel, "Self-organized growth of quantum-dot structures," *Semicond. Sci. Technol.* **11**, 1365 (1996).
17. N. N. Ledentsov, M. Grundmann, N. Kirstaedter *et al.*, "Luminescence and structural properties of (In,Ga)As—GaAs quantum dots," in: D. J. Lockwood, ed., *Proc. ICPS22* vol. 3, Singapore: World Scientific, 1995, p. 1855.
18. J.-Y. Marzin, J. M. Gerard, A. Izraël, D. Barrier, and G. Bastard, "Photoluminescence of single InAs quantum dots obtained by self-organized growth on GaAs," *Phys. Rev. Lett.* **73**, 716 (1994).
19. M. Grundmann, J. Christen, N. N. Ledentsov *et al.*, "Ultrannarrow luminescence lines from single quantum dots," *Phys. Rev. Lett.* **74**, 4043 (1995).
20. S. Nakamura, M. Senoh, S. Nagahama *et al.*, "Subband emissions of InGaN multi-quantum-well laser diodes under room-temperature continuous wave operation," *Appl. Phys. Lett.* **70**, 2753 (1997).
21. J. M. Gérard, O. Cabrol, and B. Sermage, "InAs quantum boxes: highly efficient radiative traps for light emitting devices on Si," *Appl. Phys. Lett.* **68**, 3123 (1996).
22. N. N. Ledentsov, J. Böhrer, D. Bimberg *et al.*, "3D arrays of quantum dots for laser applications," in: R. J. Shul, S. J. Pearton, F. Ren, and C.-S. Wu, eds, *Mat. Res. Soc. Symp. Proc.* vol. 421, Pittsburgh, 1996, p. 133.
23. J. Temmyo, E. Kuramochi, M. Sugo *et al.*, "Quantum disk lasers with self-organized dot-like active regions," in: *Proc. Laser Electro-Opt. Soc.* vol. 1, 1995, pp. 77-78.
24. K. Kamath, P. Bhattacharya, T. Sosnowski, and J. Phillips, "Room temperature operation of $\text{In}_{0.4}\text{Ga}_{0.6}\text{As}$ self-organized quantum dot lasers," *Electron. Lett.* **30**, 1374 (1996).
25. H. Shoji, K. Mukai, N. Ohtsuka, M. Sugawara, T. Uchida and H. Ishikawa, "Lasing at three-dimensionally quantum-confined sublevels of self-organized $\text{In}_{0.5}\text{Ga}_{0.5}\text{As}$ quantum dots by current injection," *IEEE Photon. Technol. Lett.* **7**, 1385 (1995).
26. F. Heinrichsdorff, M.-H. Mao, A. Krost *et al.*, "Room temperature lasing from vertically stacked InAs/GaAs quantum dots grown by metalorganic chemical vapor deposition," *Appl. Phys. Lett.* **71**, 22 (1997).
27. N. Kirstaedter, O. G. Schmidt, N. N. Ledentsov *et al.*, "Gain and differential gain of single layer InAs/GaAs quantum dot injection lasers," *Appl. Phys. Lett.* **69**, 1226 (1996).
28. M. V. Maximov, Yu. M. Shernyakov, A. F. Tsatsul'nikov *et al.*, "High-power continuous-wave operation of a InGaAs/AlGaAs quantum dot laser," *J. Appl. Phys.* **83**, 5561 (1998).
29. V. M. Ustinov, A. R. Kovsh, A. E. Zhukov *et al.*, "Low threshold heterostructure laser based on quantum dots emitting at $1.84\text{ }\mu\text{m}$," *Pis'ma Zh. Tekh. Fiz. (Tech. Phys. Lett.)* **24** (1), 49 (1998).
30. A. Moritz, R. Wirth, A. Hangleiter, A. Kurtenbach, and K. Eberl, "Optical gain and lasing in self-organized InP/GaInP quantum dots," *Appl. Phys. Lett.* **69**, 212 (1996).

31. D.L. Huffaker, O. Baklenov, L. Graham, B. G. Streetman, and D. G. Deppe, "Quantum dot vertical-cavity surface-emitting laser with a dielectric aperture," *Appl. Phys. Lett.* **70**, 2356 (1997).
32. J. A. Lott, N. N. Ledentsov, V. M. Ustinov *et al.*, "Vertical cavity lasers based on vertically coupled quantum dots," *Electron. Lett.* **33**, 1150 (1997).
33. L. V. Asryan and R. A. Suris, "Inhomogeneous line broadening and the threshold current density of a semiconductor quantum dot laser," *Semicond. Sci. Technol.* **11**, 554 (1996);
L. V. Asryan and R. A. Suris, "Charge neutrality violation in quantum-dot lasers," *IEEE J. Selected Topics Quantum Electron.* **3**, 148 (1997).
34. M. Grundmann and D. Bimberg, "Gain and threshold of quantum dot lasers: theory and comparison to experiments," *Jpn. J. Appl. Phys.* **36**, 4181 (1997).
35. L. E. Vorob'ev, D. A. Firsov, V. A. Shalygin *et al.*, "Spontaneous far-IR emission accompanying transitions of charge carriers between levels of quantum dots," *JETP Lett.* **67**, 275 (1998).
36. G. E. Cirlin, V. N. Petrov, V. G. Dubrovsky *et al.*, "Growth of InAs quantum dots on silicon," *Pis'ma Zh. Tekh. Fiz. (Tech. Phys. Lett)* **24** (8), 10 (1998).
37. M. V. Maximov, N. N. Ledentsov, A. F. Tsatsul'nikov *et al.*, "Middle infrared emission from InAs quantum dots in a GaAs matrix," *Proc. ICPS24*, Jerusalem, 1998.
38. G. N. Aliev, A. D. Andreev, R. M. Daitsev *et al.*, "Interband magneto-optics of CdSe and optical spectra of CdSe-based quantum dots," *J. Cryst. Growth* **184/185**, 315 (1989).
39. N. N. Ledentsov, I. L. Krestnikov, M. V. Maximov *et al.*, "Ground state exciton lasing in CdSe submonolayers inserted in a ZnSe matrix," *Appl. Phys. Lett.* **69**, 1343 (1996).
40. N. N. Ledentsov, D. Bimberg, V. M. Ustinov *et al.*, "Self-adjustment of the cavity mode and the gain spectrum in vertical cavity quantum dot laser," *Semicond. Sci. Technol.*, in print (1998).
41. I. L. Krestnikov, M. V. Maximov, A. V. Sakharov *et al.*, "RT lasing and efficient optical confinement in CdSe/ZnMgSSe submonolayer superlattices," *J. Cryst. Growth* **184/185**, 545 (1998).
42. I. L. Krestnikov, P. S. Kop'ev, Zh. I. Alferov *et al.*, "Vertical arrangement and wavefunction control in sheets of two-dimensional CdSe islands in a ZnSe matrix," *Proc. ICPS24*, Jerusalem, 1998.
43. V. P. Kalosha, G. Ya. Slep'yan, S. A. Maksimenko *et al.*, "Gain spectrum splitting in quantum dot lasers" *Proc. ICPS24*, Jerusalem, 1998.
44. S. W. Jun, T.-Y. Seong, J. H. Lee and B. Lee, "Naturally formed $\text{In}_x\text{Al}_{1-x}\text{As}/\text{In}_y\text{Al}_{1-y}\text{As}$ vertical superlattices," *Appl. Phys. Lett.* **68**, 3443 (1996).
45. Zh. I. Alferov and R. F. Kazarinov, "Double heterostructure laser," Authors certificate No. 27448, Application No. 950840 filed on March 30, 1963.
46. A historic bet on the time frame when QD lasers will take over from QW devices was made by H. Stormer and N. N. Ledentsov at the Workshop, with the former betting that QD lasers will not take over for at least 5 years.

Nanostructure Self-Assembly as an Emerging Technology

James L. Merz

Dept. of Electrical Engineering, Univ. of Notre Dame, Notre Dame, IN 46556

Albert-László Barabási, Jacek K. Furdyna

Dept. of Physics, Univ. of Notre Dame, Notre Dame, IN 46556

R. Stanley Williams

Quantum Structures Research Initiative, Hewlett-Packard Laboratories, Palo Alto, CA 94304

1. Introduction

In this article a brief review will be given of the growth and properties of self-assembled quantum dots (SAQDs). The current understanding of the growth mechanisms will be described, and some of the techniques that might be used to control the growth will be discussed. Emphasis will be placed on controlling the nucleation and growth of these self-assembled quantum dots as a means to produce ordered arrays. Finally, a brief discussion will be given of possible applications of arrays of SAQDs.

2. Comments on the growth of SAQDs

Several different semiconductor systems will be considered in this paper including InAs quantum dots on GaAs, CdSe dots on ZnSe, and Ge dots grown on Si. The relationship between band gap and lattice constant is well known for a large number of compounds and Group IV semiconductors. The relative mismatch between the InAs and GaAs lattices is approximately 6.9%, slightly less than 7% for CdSe/ZnSe, and about 4% for Ge on Si. These values of the lattice mismatch are critical in determining the growth mode of SAQDs because the growth process is largely driven by strain considerations resulting from that lattice mismatch.

Within the III-V compounds, considerable work has been done on the InAs/GaAs system.^{1,2} Scanning tunneling microscope (STM) and atomic force microscope (AFM) images indicate that arrays of these dots can be grown with diameters typically of the order of 25 nm, with a size fluctuation of between five and seven percent of that diameter. Typical dot densities are of the order of 10^9 to $10^{11}/\text{cm}^2$. By observing optical emission from individual dots, a number of investigators have independently shown that these dots are indeed zero-dimensional quantum confined structures for electrons and holes. A variety of

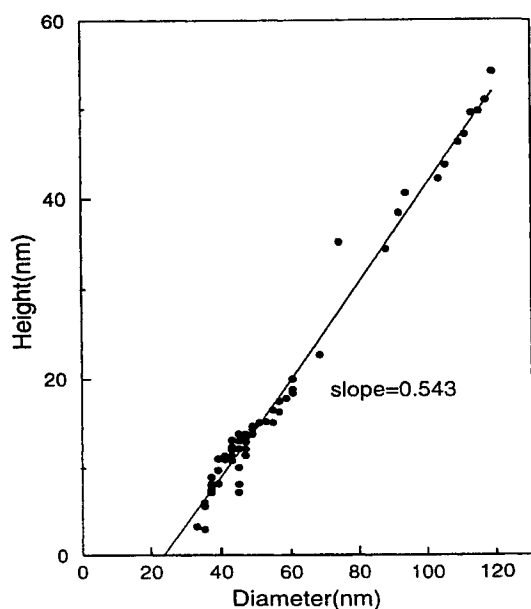


Figure 1. Dependence of height on diameter for CdSe dots on ZnSe. Note the linear dependence that extrapolates to 23nm diameter at zero height, indicating the error due to AFM tip diameter. The slope of 0.543 gives the aspect ratio.^{6,7}

techniques has been used to reduce the number of quantum dots observed in these studies from values exceeding 10^6 to just a few hundred dots, for which strong emission from single dots can be observed, showing the δ -function-like characteristic expected from a zero-dimensional structure. Examples of these experiments include luminescence excitation with an electron beam (cathodoluminescence) using small apertures over the surface of the sample, or etching away all dots except a limited number contained on mesas.

In the case of the II-VI compounds, several laboratories have been investigating the growth of CdSe on ZnSe,¹⁻⁹ and self-assembled islands are again observed. In this case the growth temperature is considerably lower than for the III-Vs, on the order of 370 °C, and typical dot sizes or islands appear to be somewhat larger, as will be discussed below. Dot densities of a few times $10^9/\text{cm}^2$ have been observed. Uncapped CdSe dots are not stable on the ZnSe surface; detailed AFM studies have shown that they undergo Ostwald ripening¹² at room temperature. Study of this ripening process has been carried out in detail.^{6,7} It has been demonstrated that all of the islands immediately after growth are of apparent size of approximately 55 nm. However, some of the dots subsequently grow in size at the expense of others. Figure 1 shows the height of a large number of these dots, measured by AFM, as a function of their diameter, regardless of the extent of ripening that has already taken place. All of the data points fall on a straight line

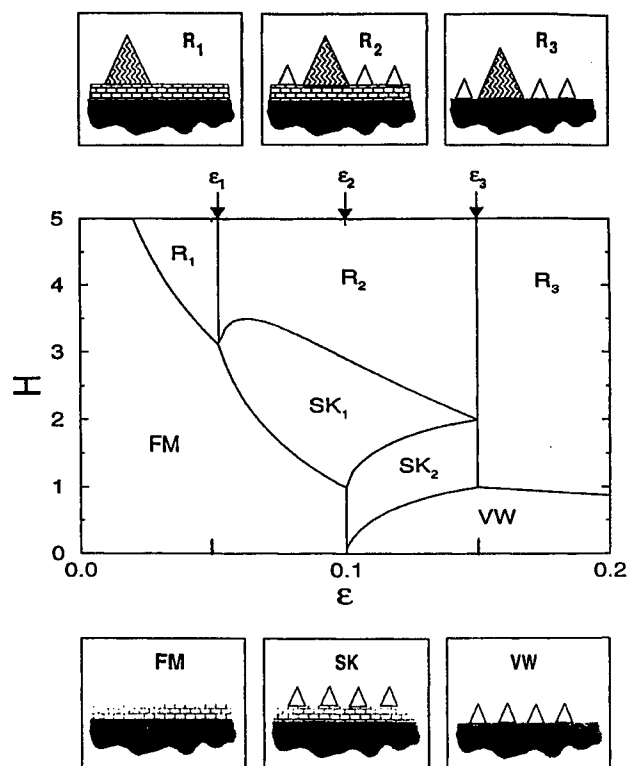


Figure 2. Equilibrium phase diagram for SAQD formation in function of the coverage (H) and lattice mismatch (ϵ). The small panels on the top and the bottom illustrate the morphology of the surface in the six growth modes. The small empty islands indicate the presence of stable islands, while the large shaded ones refer to ripened islands.¹³

that extrapolates at zero height to a diameter of approximately 23 nm. The fact that this line does not go through zero is a result of the finite size of the tip of the AFM machine. The dot diameter immediately after growth is therefore approximately $55 - 23 \approx 32$ nm. Interestingly, the dots are highly stable at 0 °C.

In order to understand the growth mechanism, an equilibrium study of dislocation-free island formation has been carried out by Daruka and Barabási.^{13,14} The results are summarized in the phase diagram shown in Fig. 2, which presents the possible growth modes in functional relationship to the number of monolayers deposited (H) and the lattice mismatch (ϵ). The phase diagram distinguishes a number of different growth mechanisms, including two-dimensional (2D) growth (or layer-by-layer growth), various kinds of Ostwald ripening indicated by R_I ($I = 1$ to 3), Stranski-Krastanow (SK), and Volmer-Weber (VW) growth. The 2D, SK,

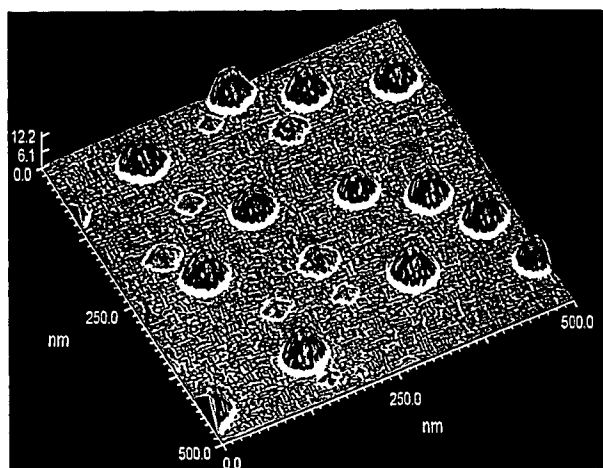


Figure 3. In situ scanning tunneling microscope image of 8 equivalent monolayers of Ge deposited onto Si(001) at 600 °C by physical vapor deposition, which shows the coexistence of both Ge pyramids and domes on the surface. In this image, the different facets are keyed to different shades of gray and the edges are enhanced by including a component of the local Laplacian in the shading of the image.

or VW growth modes generate stable islands, whereas three different kinds of ripening which may or may not include a two-dimensional layer (frequently called the wetting layer) are possible. This diagram demonstrates that in order to achieve stable and controllable SK or VW growth, for which uniform small islands may result, it is necessary to employ semiconducting crystal systems that include a large amount of strain (i.e., work far out along the horizontal axis in this diagram), and the deposited material must not exceed a certain critical coverage. For the systems already described (InAs/GaAs and CdSe/ZnSe), it appears that growth has taken place in the region $0.05 < \epsilon < 0.1$, resulting either in a 2D-to-ripening transition for CdSe on ZnSe,^{6,7} or possibly SK growth for InAs on GaAs.

Because of its compatibility with silicon microelectronic VLSI, a more interesting system is Ge on Si. Typically,¹⁵ an array of dots is observed with average heights of 15 nm, a standard deviation less than 1 nm and a dot density of $6 \times 10^9 \text{ cm}^{-2}$. However two different kinds of islands are usually seen, as shown in Fig. 3. If 8 monolayers of Ge are deposited on Si by physical vapor deposition (e.g., MBE), both small pyramidal islands, with the base of the pyramid of order 20 to 25 nm, and larger dome-shaped islands that exceed 50 nm are observed.¹⁶ If growth is stopped after only four monolayers of Ge, most of the islands are the small pyramid type. The distribution of island sizes is shown in Fig. 4, where the distributions of smaller pyramids and larger domes are shown as a function of island volume. Both of these island types (pyramids and domes) are larger than

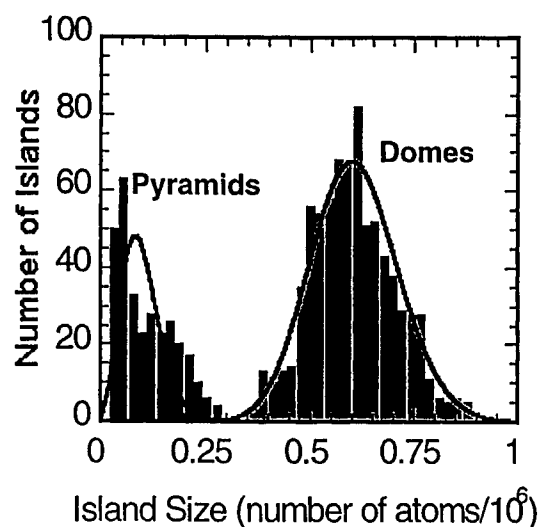


Figure 4. The histograms are the experimentally measured volume distributions of the pyramids and domes measured from high-resolution STM topographs of the same sample as Figure 5. The solid lines are the fits to the experimental data using the theory of Shchukin *et al.*¹⁷

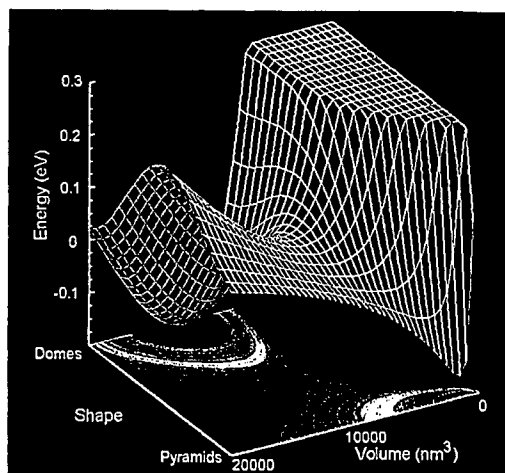


Figure 5. Internal free energy per atom vs. size and shape for Ge nanocrystals on Si(001). The volume dependence for the pyramids and domes was determined from the fit of Fig. 4. The shape axis is the reaction coordinate Γ taking a pyramid ($\Gamma = 0$) to a dome ($\Gamma = \pi$). The functional form for the plot along the shape axis is $E_a \sin^2[\Gamma]$ added to a linear interpolation between the energies of the limiting shapes. The saddle point represents the transition state of the shape change.

desired for quantum-confined systems to be of use for electronic applications, but considerable information about growth mechanisms has resulted from this systematic study.

An experimental STM study of the mechanism of Ge growth on Si showed that the size distributions of the pyramids and domes was consistent with a Boltzmann distribution applied to the free energy prediction of Shchukin *et al.*¹⁷ A fit to experimental data, including the activation barrier for the shape transition, yields the free energy plot shown in Fig. 5; this plot reveals energy minima for both pyramids and domes as a function of the volume of the islands, and the activation barrier as a function of the shape parameter.

3. Nucleation control

In order for SAQDs to be useful from a practical point of view, it is essential to control the nucleation of these islands, because an ordered array of small dots that are equal in size is required for most applications. The size of the dots can be controlled either by strain resulting from the lattice mismatch of the system, as described above, or by limiting the source of atoms. For example, one can limit the surface area from which atoms are available for the growth of a single dot by fixing nucleation centers with spacing of the order of the surface diffusion length of these atoms. One can also order the array of resulting dots by some sort of coded or directed assembly procedure, by which nucleation sites are produced through some sort of surface patterning process. An example is shown in Fig. 6.

The upper part of this figure shows six simulated dot arrays, three of which result when growth is initiated on an unpatterned substrate, and three for which the substrate has been pre-patterned.¹⁸ Three different values of growth flux have been used, $F = 0.03, 0.08, 0.3$ ml/s, for which the diffusion length of surface atoms (l_d) is taken to be greater than, approximately equal to, or less than the pattern spacing, respectively. (Note that l_d decreases with increasing F , and increases with increasing growth temperature).¹⁹ The fluctuation in island size resulting in the case of substrate patterning is shown by the curve at the bottom of Fig. 6, which plots the normalized width of the dot size distribution as a function of flux. The square points indicate the case of no patterning. These simulations show that if the substrate is patterned with ordered nucleation sites separated by a distance l_i , then the fluctuation in island sizes varies significantly, depending on the size of the dot spacing relative to the diffusion length, with a sharp minimum at $l_d \approx l_i$. Thus, if the substrate can be patterned with spacing comparable to the diffusion length, regularly spaced islands with a small and uniform size can be obtained. Since the diffusion length can be controlled with the temperature and flux, one can find an appropriate combination of T and F to obtain optimal growth conditions for any desired island size.

How can one pattern the substrate before growth? Two techniques have been suggested, (1) growth of narrow mesas on the substrate, so that SAQDs can form only on the top of a mesa that extends above a passivating oxide,²⁰ and (2) using

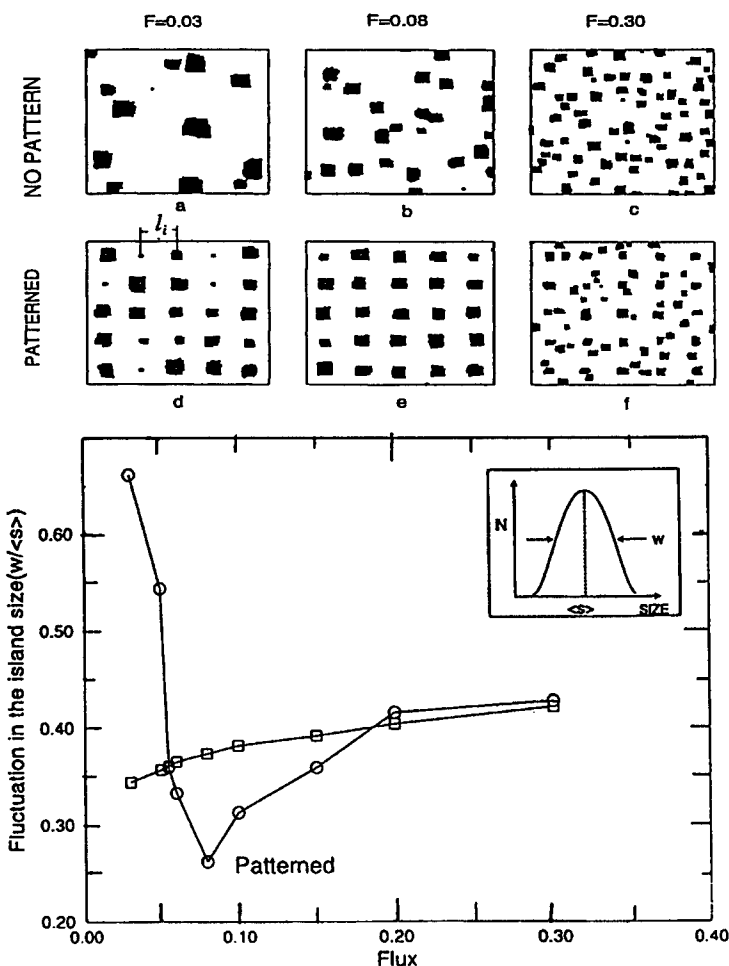


Figure 6. Top: Island morphologies obtained on an impurity-free substrate (a-c) as well as on a surface patterned with impurities (d-f) for coverage $\theta = 0.1$. The parameters in the Monte Carlo simulations are system size $L = 400$, $T = 400$ K, and $l_i = 40$. Bottom: Relative width of the island size distribution for surfaces without (A) and with impurities (B). $W/\langle s \rangle$ increases monotonously for the surface without impurities (A). In the presence of impurities, however, there exist a global minimum around $F_{\text{opt}} \approx 0.08$.¹⁸

scanning probe techniques, such as STM or AFM, to produce nucleation sites. In the first case, nucleation sites tend to form equally-spaced along the mesa, because the formation of the mesa itself provides equal areas to serve as the source of atoms, as shown in Fig. 7. Two rows of regularly-spaced quantum dots are observed on each edge of the mesa. If the width of the mesa is increased, for

example to 670 nm or 1.0 micron, then additional rows of dots are formed, and the spacing of the dots is randomized, particularly for those that grow in the center of the mesa.²⁰

An alternate technique for fixing nucleation sites employs scanning probe techniques, such as STM or AFM patterning. With the use of these scanning probes, one could (a) add foreign atoms, (b) remove host atoms, or (c) produce point defects or clusters, each of these serving as potential nucleation centers. The problem with this technique is that of *speed*; for a large array of dots, e.g., as might be needed for a modern VLSI architecture, it could take weeks, months, or even years to produce such a pattern by scanning probe techniques. However the speed by which scanning probes can be moved has been steadily increasing, with writing speeds up to 10 $\mu\text{m/s}$ reported recently.²¹ Also, there is ongoing research aimed at producing large numbers of parallel tips that can be controlled simultaneously.^{22,23} The current state of the art involves 144 tips moving together. Any of these tips could be activated electrically, so that it should be possible to produce a pattern that is coded in a way to produce the desired circuit architecture. In general, it is possible that large-scale patterning by lithography combined with short-length-scale patterning by STM or AFM will provide the speed and reliability needed for applications such as quantum computers, lasers or detectors.

4. Applications

An important application for ordered arrays of SAQDs is the so-called quantum cellular automata (QCA) scheme developed by Lent and co-workers.²⁴ This

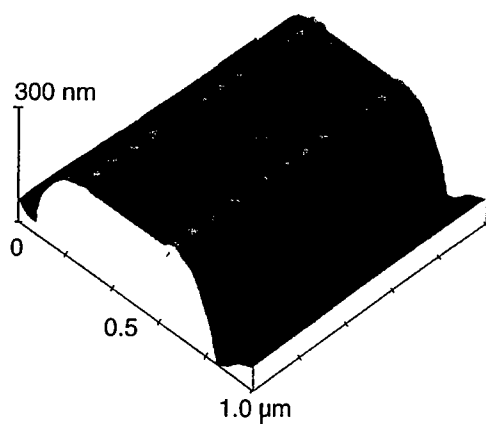


Figure 7. An atomic force microscope image of a 450 nm wide Si mesa oriented along a [100] axis of a (001) surface. The mesa was formed by growing Si on a surface that was patterned with a silicon dioxide film to restrict the regions where Si could grow. After formation of the mesa, Ge was deposited by chemical vapor deposition using germane. The Ge islands formed preferentially along the edges of the mesa.

application features arrays of quantum cells, each of which consists of either 4 or 5 quantum dots, with four dots occupying the four corners of a square. The cell is populated with two electrons that can tunnel between dots in a given cell, but cannot escape from the cell. The two possible ground states of this system would then be the two possible configurations of electrons opposite each other at the corners of the cell, as determined strictly by classical Coulomb repulsion. If one then fixes the configuration of one cell, the configuration of the neighboring cell is also fixed by Coulomb repulsion. It can be shown that the response function of this system, i.e. the response of one cell to its neighbor, is highly non-linear, which represents a kind of gain for the system. This nonlinearity permits the robust encoding of binary information in the state of the cell. Using this system, it has also been shown that all of the necessary logic operations for computing can be performed, and that this can be done in a highly efficient arrangement in terms of chip area.^{25,26} For example, a full adder can be designed with approximately 200 of these cells (800 quantum dots) in a space less than $1.5 \mu\text{m}^2$, whereas a full adder using conventional silicon CMOS technology would utilize 30 transistors at an average packing density of about $1.1 \mu\text{m}^2/\text{transistor}$ (using the SIA National Semiconductor road map for the year 2010). Furthermore, the highly complex multilevel metallization procedures now required for the wiring of conventional integrated circuits would be obviated because the QCA concept is truly a cellular automata system, such that contacts need only be made to an input plane and output plane—information propagates through the architecture in a fully automatic fashion.

Using conventional semiconductor quantum dots of the sort discussed earlier in this paper, for which typical nearest neighbor distances of these dots is of order 20 nm, the highly nonlinear switching function described above requires low-temperature operation (4 K). However, a molecular electronics implementation of this scheme is possible, for which the nearest neighbor distances are approximately 2 nm. In this case, the operational temperature of the system can increase by two orders of magnitude; i.e., operating temperatures well above room temperature are possible. A possible molecular implementation involves use of metal cluster carboxylates, as proposed by Cen *et al.*,²⁷ for which identical clusters involving the transition metals form square cells of the sort required. Research along these lines is currently proceeding at several laboratories.

5. Conclusions

In this paper we have described how the understanding of growth mechanisms of self assembled quantum dots has increased in recent years, and we described the current status for three semiconducting systems: (1) InAs/GaAs (III-V) technology, for which stable and uniform dots can be obtained, (2) CdSe/ZnSe (II-VI) dots, which show complicated ripening or coarsening phenomena, even at room temperature, and (3) Ge/Si (group IV) structures, whose growth shows a high degree of control, but the resulting dots are too large for quantum-

confinement purposes. We have described new methods for controlling uniformity that are being explored as well as methods to control (or code) the nucleation events. Finally, an application involving quantum cellular automata that could utilize semiconductors (if operated at low temperatures) or molecular-level structures (which operate at room temperature) was briefly described.

6. Acknowledgments

The authors wish to thank C. Lent for discussions on QCA, C.-S. Lee for providing Fig. 1, and T. I. Kamins for Fig. 7. The authors wish to thank the following agencies for support: J. L. M. was supported by DARPA/ONR (N00014-95-1-1166); A.-L. B. was supported by the ONR Young Investigator Program (N00014-98-1-0575); and J. K. F. was supported by the NSF (DMR 97-05064).

References

1. For up-to-date reviews see, e.g., W. Seifert, N. Carlsson, M. Miller, *et al.*, "In-situ growth of quantum dot structures by the Stranski-Krastanow growth mode," *J. Crystal Growth Character. Mater.* **33**, 423 (1996) and R. Nötzel, "Self-organized growth of quantum-dot structures," *Semicond. Sci. Technol.* **11**, 1365 (1996).
2. P. M. Petroff and G. Medeiros-Ribeiro, "Three dimensional carrier confinement in strain-induced self-assembled quantum dots," *MRS Bull.* **21** (4), 50 (1996).
3. S. H. Xin, P. D. Wang, Aie Yin, *et al.*, "Formation of self-assembling CdSe quantum dots on ZnSe by molecular beam epitaxy," *Appl. Phys. Lett.* **69**, 3884 (1996).
4. F. Flack, V. Nikitin, P. A. Crowell, *et al.*, "Near-field optical spectroscopy of localized excitons in strained CdSe quantum dots," *Phys. Rev. B* **54**, R17312 (1996); see also F. S. Flack, A. Hunt, H. Hennessey, *et al.*, "Growth dynamics and exciton localization in strained CdSe quantum structures," *Mat. Res. Symp. Proc.* Vol. 417, Pittsburgh: Materials Research Society, 1996, p. 169.
5. J. Merz, S. Lee, and J. K. Furdyna, "Self-organized growth, ripening, and optical properties of wide-bandgap II-VI quantum dots," *J. Crystal Growth* **184/185**, 228 (1998).
6. J. K. Furdyna, S. Lee, I. Daruka, *et al.*, "Self-assembled growth of II-VI quantum dots," *Nonlinear Opt.* **18**, 85 (1997).
7. J. K. Furdyna, S. Lee, A.-L. Barabási, and J. L. Merz, "Self-organized low-dimensional II-VI nanostructures," to appear in: M. Tamargo, ed., *II-VI Semiconductor Materials and Their Application*, New York: Gordon and Breach, 1999.
8. K. Leonardi, H. Heinke, K. Ohdawa, *et al.*, "CdSe/ZnSe quantum structures grown by migration enhanced epitaxy: structural and optical investigations," *Appl. Phys. Lett.* **71**, 1510 (1997).

9. D. Hommel, K. Leonardi, H. Heinke, *et al.*, "CdSe/ZnSe quantum dot structures: structural and optical investigations," *Phys. Stat. Sol. (b)* **202**, 835 (1997).
10. E. Kurtz, H. D. Jung, T. Hanada, *et al.*, "The growth and photoluminescence properties of self-organized CdSe quantum dots on a (111)A ZnSe Surface," *Nonlinear Opt.* **18**, 13 (1997).
11. E. Kurtz, H. D. Jung, T. Hanada, *et al.*, "Self-organized CdSe/ZnSe quantum dots on a ZnSe (111)A surface," *J. Crystal Growth* **184/185**, 248 (1998).
12. W. Ostwald, *Z. Phys. Chem. (Leipzig)* **34**, 495 (1990); for a review, see M. Zinke-Allmang, L. C. Feldman, and M. H. Grabow, "Clustering on surfaces," *Surf. Sci. Rep.* **16**, 377 (1992).
13. I. Daruka and A.-L. Barabási, "Dislocation-free island formation in hetero-epitaxial growth: a study at equilibrium," *Phys. Rev. Lett.* **79**, 3708 (1997).
14. I. Daruka and A.-L. Barabási, "Equilibrium phase diagrams for dislocation free self-assembled quantum dots," *Appl. Phys. Lett.* **72**, 2102 (1998).
15. T. I. Kamins, E. C. Carr, R. S. Williams, and S. J. Rosner, "Deposition of three-dimensional Ge island on Si(001) by chemical vapor deposition at atmospheric and reduced pressures," *J. Appl. Phys.* **81**, 211 (1997).
16. G. Medeiros-Ribeiro, A. M. Bratkovski, T. I. Kamins, D. A. A. Ohlberg, and R. S. Williams, "Shape transition of germanium nanocrystals on a Si(001) surface from pyramids to domes," *Science* **279**, 353 (1998).
17. V. A. Shchukin, N. N. Ledentsov, P. S. Kop'ev, and D. Bimberg, "Spontaneous ordering of arrays of coherent strained islands," *Phys. Rev. Lett.* **75**, 2968 (1995).
18. C.-S. Lee and A.-L. Barabási, "Spatial ordering of islands grown on patterned surfaces," submitted to *Appl. Phys. Lett.* (1998).
19. A.-L. Barabási and H. E. Stanley, *Fractal Concepts in Surface Growth*, Cambridge, U.K.: Cambridge University Press, 1995.
20. T. I. Kamins and R. S. Williams, "Lithographic positioning of self-assembled Ge islands on Si(001)," *Appl. Phys. Lett.* **71**, 1201 (1997).
21. S. W. Park, H. T. Soh, C. F. Quate, and S.-I. Park, "Nanometer scale lithography at high scanning speeds with the atomic force microscope using spin on glass," *Appl. Phys. Lett.* **67**, 2415 (1995).
22. R. F. Service, "Meeting briefs: atomic landscapes beckon researchers," *Science* **274**, 723 (1996).
23. S. C. Minne, S. R. Manalis, and C. F. Quate, "Parallel atomic force microscopy using cantilevers with integrated piezoresistive sensors and integrated piezoelectric actuators," *Appl. Phys. Lett.* **67**, 3918 (1995).
24. Craig S. Lent, P. Douglas Tougaw, Wolfgang Porod, and Gary H. Bernstein, "Quantum cellular automata," *Nanotechnology* **4**, 49 (1993).
25. P. Douglas Tougaw and Craig S. Lent, "Logical devices implemented using quantum cellular automata," *J. Appl. Phys.* **75**, 1818 (1994).
26. Craig S. Lent and P. Douglas Tougaw, "A device architecture for computing with quantum dots," *Proc. IEEE* **85**, 541 (1997).
27. W. Cen, K. J. Haller, and T. P. Fehlner, "On the role of PES data in the identification of metal-metal charge transfer bands in clusters of clusters," *J. Electron Spectrosc.* **66**, 29 (1993).

Nonlithographic Fabrication and Collective Behavior for Future Nanoelectronics and Computation

A. J. Bennett, D. Levner, J. Li, C. Papadopoulos, A. Rakitin, and J. M. Xu

Optoelectronics and Emerging Technologies Laboratory, Dept. of Electrical and Computer Engineering, University of Toronto, Toronto, Canada, M5S 3G4

1. Introduction

The great success of microelectronics has not been due to radical technological revisions during its evolution, but due to the remarkable performance improvements enabled by miniaturization. Few, if any, fundamental changes have been made to the planar process of Noyce from 1958 and the MOS transistor scaling ideas proposed in the early 1970's¹ — they have instead been refined to an unprecedented level. Despite its undeniable success, the microelectronics industry is reaching a crossroads. Previously routine semiconductor process improvements now demand serious or even heroic R&D and investment, and fabrication plants now cost billions of dollars to build.² In fact, marching in parallel with the famed Moore's Law, which correctly predicted the doubling of semiconductor system performance every 18 months, has been a less discussed but equally important law on the economic side: the capital investment for the equipment and research required to fuel Moore's Law also doubles every generation.³

This economic law suggests that even if basic physical and technological problems are surmountable, the semiconductor industry will be first forced to address the exploding costs of research and equipment and to balance them against the diminishing benefits realized by following the one-dimensional track of miniaturization. As costs increase exponentially while return-on-investment in hardware technology decreases, all but the largest players will be squeezed out, and capital will flow to the areas in which radical innovation and significant performance improvement are still possible: e.g. novel silicon architectures, system designs, and software. This capital flow will consequently slow down innovation in hardware technologies and lead to a plateau in Moore's Law.

Thinking optimistically, the saturation of Moore's Law may still be a blessing in disguise. Since no one can predict the final outcome as hard economic and technical limits are reached over the next two decades, the future of computing need not be determined exclusively by the major players in the industry, or others with billions of dollars to invest. Instead, one can anticipate an explosion in creativity — new device, architecture and system designs — that will be required to address the burgeoning problems arising from miniaturization, interconnection, and heat dissipation. The search for alternative paths to nanoelectronics opens up the future to many more potential players, and likely new ones as well.

Alternatives to the well-trodden path of miniaturization do exist. Nonlithographic nanofabrication is one example—a particularly attractive one given that lithography is the most costly part of the current technology. It is becoming increasingly apparent that "natural" approaches — e.g. self-organization and/or self-assembly — can produce regular, well-ordered, high quality structures on the scale of nanometers. In such fabrication processes, physical and chemical laws, acting and/or competing on small characteristic length scales, conspire to build or "grow" structures at the nanometer scale. Such approaches typically require relatively little human intervention in comparison to photo or electron-beam lithography.

Indeed, it is possible that while circumventing some of the difficulties faced by lithography, non-lithographic nanofabrication techniques can also address some of the most significant problems facing current and near-future microelectronic systems: wiring, heat dissipation, and quantum effects.

In order to keep the resistance of circuit interconnects low, it is desirable to make them as broad as possible. However, this requirement contradicts the philosophy of miniaturization, which improves the functionality and power of semiconductor systems and devices by increasing packing density. While research laboratories may achieve the routine production of sub-100 nm minimum feature sizes,⁴ interconnecting these devices to form useful multi-million transistor circuits is a serious challenge with no immediate solution. Indeed, we are on the verge of observing fascinating but undesirable physical effects such as conductance quantization and ballistic transport in VLSI interconnects. Such wires are no longer passive lines of connection, but instead participate in nonlinear processing of digital signals, and should be regarded more as functional devices, thus rendering invalid current circuit design methodologies based on existing connectivity theory.

Moreover, the total wiring length in modern microprocessors is now on the order of kilometers per square centimeter. While such numbers are a testament to impressive engineering, it is not clear how much they can be increased, nor what the associated performance improvements or trade-offs will be. Advances such as IBM's highly publicized copper interconnect process,⁵ while important on the scale of years and months, represent one-time, evolutionary changes to an existing technology, and not the revolutionary changes which will be required to guide microelectronics through the next few decades.

With today's fastest processors dissipating approximately 30 W/cm², heat dissipation will become an even greater problem as packing density continues to rise.⁶ There is certainly room for improvement, since microprocessors dissipate approximately 10⁻⁶ J per bit switched, while the human brain is estimated to dissipate only 10⁻¹⁶ J per bit.⁷ In view of these numbers, it will be less and less viable to charge up increasingly narrow and resistive transmission lines to communicate between logic blocks. "Wireless" computing alternatives, such as those that depend on electronic polarization of nanostructures, may have an advantage in solving this problem, since little conduction current — and hence almost no Joule heating — is associated with signal transmission.⁸

Yet another looming problem that will affect future generations of deep-submicron circuits is the failure of the fundamental assumptions underlying the semi-classical Boltzmann transport equation—a direct consequence of the success of miniaturization.⁹ As device dimensions drop below the dephasing length and/or the mean free path of the electron (>100 nm), we enter the realm of quantum transport. The phase of the electron wavefunction becomes relevant, and phenomena such as ballistic transport, coherent interference, and reflection have observable effects. At still smaller size scales (~ 1 – 10 nm), quantum tunneling becomes significant. All of these effects disrupt such basic digital design rules as signal cascadability, fan-out, binary logic restoration, and input-output isolation.

As minimum feature sizes continue to decrease, fabrication reliability also worsens, since statistical averaging over decreasing numbers of dopants implies increasing relative deviations. Therefore, as device counts on a chip rise, the performance deviations scale up inversely to device volume, dramatically reducing the yield of defect-free chips. Indeed it may no longer be technically or economically feasible to demand defect-free chips. The hardware fault-tolerant "Teramac" project at Hewlett-Packard¹⁰ is an example of one massively parallel computer architecture initiative intended to address these problems. Such a system demonstrates that a properly designed architecture can circumvent the imperfect yield of devices, and suggests defects in self-organized nanosystems need not interfere with their potential utility as computational machines.

We envision a goal of computing using non-lithographic and/or self-organized nanostructures, although not by attempting to implement VLSI on a smaller scale with new nanodevices — i.e. not by simply continuing to exploit the heretofore successful paradigm of miniaturization. The advanced state of the art in VLSI design and devices, and the problems presented by mesoscopic physical effects suggest that a radically different design approach will be required to realize any advantage over existing technologies and methods. Another more fundamental reason for employing a different approach is that we give ourselves the opportunity to exploit directly the physics of nanostructures. Consequently, any study of potential applications of coupled nano-arrays must begin with the basic physics of electronic transport in the nanostructures, and with the physics of interactions between individual elements. In fact, we will argue that the symmetric and long-range coupling between nanostructures in an array raises the possibility of engineering a neural network or cellular automaton-like system to solve hard computational problems. Such a system would encode a massively parallel computer directly in a physical system, and would compute through the interactions between elements, circumventing the problem of interconnection and possibly that of the power dissipation associated with binary switching.

Despite such fascinating possibilities, there exist numerous challenges to the development of systems based on non-lithographic nanostructures. Present day nanostructured materials tend either to be randomly patterned, or ordered but in too small a number, or else to have a very simple lattice structure repeated many millions of times — neither a random (amorphous) lattice nor a regular lattice

replicates the engineered complexity and information content of a VLSI microprocessor. Therefore, much challenging work lies ahead.

The organization of the paper is as follows: in Section 2 we survey a number of techniques of non-lithographic fabrication of nanostructures, but principally our preferred method of nano-fabrication using an anodic aluminum oxide (AAO) template; in Section 3 we describe some of the novel physics predicted to occur in close packed nano-arrays and motivate future device and system applications; in Section 4 we describe our nanostructures in the context of potential neural computation systems; finally, in Section 5 we present our conclusions.

2. Survey of nanofabrication techniques

The principal methods for fabricating nanomaterials developed so far include conventional photolithography-based techniques,¹¹ direct-write electron-beam lithography,¹² self-assembling or self-organizing nanoparticles or quantum dots,¹³ sol-gel processing,¹⁴ nanochannel glass (NCG) templates,¹⁵ electroplating of polycarbonate membranes,¹⁶ and anodic aluminium oxide (AAO) templates.¹⁷

The fabrication of structures with feature sizes smaller than 100 nm is an expensive and technically daunting task using conventional (e-beam and photo) lithographic methods. Therefore, it is natural to seek other possible fabrication techniques — nonlithographic methods — that may be better suited for mass production of nanoscale devices. Most approaches fall into one of three main categories: (i) template methods; (ii) self-organized nanostructures; and (iii) direct growth. In this discussion, we will focus primarily on the nanotemplate approach using anodic alumina, but will describe briefly a number of different approaches. We will also mention some hybrid methods that incorporate a lithographic step into their process.

- *Formation of nanotemplate and nanostructure arrays*

Porous aluminium oxide films formed by electrochemical anodization processes in various electrolytes have been studied for more than 40 years.¹⁸ This technique has been widely used by industry for aluminium surface coloring. Anodic porous alumina can be fabricated to exhibit a packed array of columnar hexagonal cells with cylindrical, uniformly sized pores 4–200 nm in diameter. The diameter and length of the pore may be controlled by selecting the electrolyte composition and concentration, and the anodization voltage and time. The main mechanism leading to straight and uniform pore formation is believed to be the field-assisted dissolution of the oxide.¹⁹

This technique was recently observed by Masuda *et al.* to produce a highly ordered, hexagonal close-packed "honeycomb" structure under appropriate processing conditions.²⁰ The mechanism of forming the highly ordered self-organized structure is not yet fully understood.^{21,22} Figure 1 shows SEM and AFM images of a typical example of the self-organized anodic alumina template with highly ordered pores, recently fabricated in our laboratory.

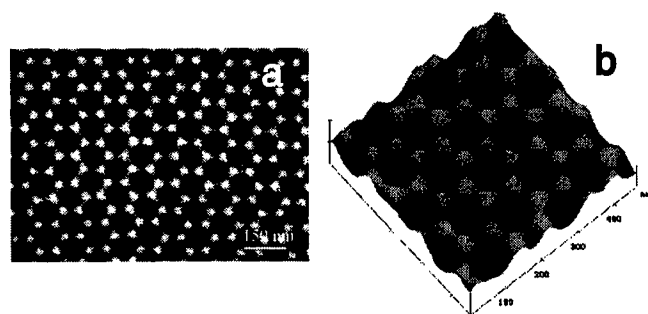


Figure 1. (a) SEM image of ordered AAO nanotemplate exhibiting hexagonal close packed lattice. (b) AFM close-up of nanotemplate showing topography.

After fabricating the AAO template, nanostructures may be grown in the pores by a number of different methods: (i) electrochemical deposition of metals, alloys and compounds into the pores of the template,²³ (ii) electrophoretic filling of the pores with colloids,²⁴ (iii) filling with sol gel via capillary action or with metals at high pressure²⁵; (iv) filling through CVD or polymerization.^{26,27}

Electrochemical deposition of metals in the AAO template has been used to produce nano-wire arrays with interesting magnetic and optical properties, differing from those of the bulk materials.²³ Li *et al.* developed a method for first producing and subsequently filling arrays of carbon nanotubes in the AAO template — see Fig. 2.²⁶ The desired interior metal, such as nickel or cobalt, was deposited within the nanotubes by electroless deposition.

The AAO template nanofabrication method possesses many advantages. It provides a simple and inexpensive way to fabricate large area, highly ordered, high density (10^{11} cm^{-2}) arrays of close packed nanopores that can be filled with wires, dots and even tubes in a large range of materials. The potential applications for such a technology are numerous, and include ultra-high density magnetic storage, field-emission displays, and optical/infra-red detectors. These possibilities have been described in the literature.²⁸ The recent progress in forming highly ordered and uniform array structures makes these possibilities all the more promising.

- *Other non-lithographic nanofabrication techniques*

There exist many other nonlithographic nanofabrication techniques besides AAO. Other template based techniques include: track etch membranes, in which polycarbonate or polyester templates are bombarded with fission products and etched to create channels in the damaged regions; mesoporous or molecular templates such as zeolites and molecular sieves,²⁹ and nanochannel glass arrays.³⁰

Self-organized growth of nanostructures is another popular approach, principally via the Stranski-Krastanow semiconductor growth mode³¹ in which nanoscale islands form during the growth of highly strained layers. Finally, direct-write nanofabrication methods have been developed using scanning probe microscopy,³² and atom lithography employing the standing wave pattern of a laser.³³

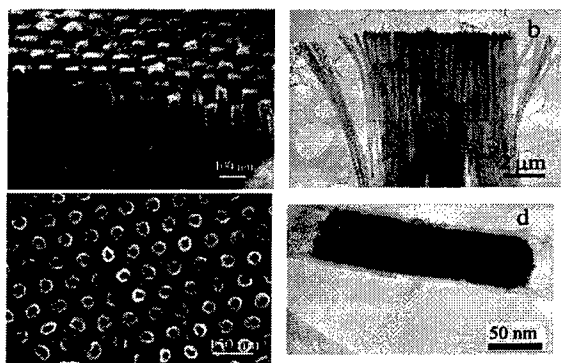


Figure 2. (a) SEM image of carbon nanotubes partially exposed by etching the alumina template with NaOH; (b) TEM image of free carbon nanotube bundle after long etch; (c) SEM top view of exposed nanotubes; (d) TEM image of a P-doped Ni particle deposited in the interior of a carbon nanotube.

Many of the methods outlined above can be used in conjunction with conventional lithography. For example, templates and self-assembled nanostructures can be used as lithographic masks. Additionally, the mechanical patterning of substrates using nanotemplates formed via electron-beam lithography prior to self-assembly or anodization can lead to a much higher degree of ordering. It is desirable in these cases that any serial lithographic step be used only once. A hybrid method of microcontact printing was used by Whitesides *et al.* to transfer self-assembled monolayer patterns onto a substrate by etching.³⁴ A unique feature of this method is its ability to produce nanostructures on curved surfaces. Using a similar technique Pang *et al.* have been able to press a wide variety of nanopatterns onto the surface of aluminum.³⁵

3. Cooperative phenomena and collective excitations in nanosystems

A number of surprising and interesting phenomena arise in the study of nanostructure interactions and collective behavior. In this section we discuss some theoretical investigations of close-packed nano-arrays that are motivated by the prospect of novel computational architectures and/or methodologies.

On the nanoscale, statistical approximations commonly used in macroscopic physics lose validity as the "granularity" of atoms and electrons becomes apparent. As a result, such phenomena as electronic screening, electronic band formation, electron-phonon interaction, *etc.* exhibit peculiarities which are absent in "orthodox" solid state physics. A remarkable example are carbon nanotubes. These nanostructures exhibit a drastic modification of the band structure with a simple change of the nanotube diameter,³⁶ with single-walled carbon nanotubes

exhibiting both semiconductor and metallic behavior depending on helicity,^{37,38} and multi-walled nanotubes behaving like amorphous carbon.³⁹

One can anticipate the appearance of nontrivial cooperative phenomena in a periodic array of nanostructures.⁴⁰ Spontaneous polarization in 2D arrays of quantum dots⁴¹ and double-dot quantum molecules⁴² was predicted to yield a ferroelectric transition. Cooperative behavior of such arrays of quantum dots or wires is potentially useful for computational hardware akin to cellular automata and neural networks.⁴³ For example, consider an array of vertically-oriented quantum wires placed between top and bottom electrodes. In individual wires, the Coulomb interaction alters the charge density distribution from that given by the conventional Luttinger liquid model.⁴⁴ The charge density rapidly decays from the ends to the middle of the wire, with an exponential cutoff at $l \approx (a/e)(\epsilon v_F)$ (where $\hbar = 1$, a is the wire lattice parameter, v_F is the Fermi velocity, and ϵ is the dielectric constant of the medium surrounding the wires). Thus, if the interwire distance in the array is sufficiently large compared to the wire length ($d \gg L$), we have an interacting array of dipoles. The peculiarity of this system lies in the competition between two tendencies: the external bias tends to maintain ferroelectric order whereas the dipole-dipole interaction promotes the antiferroelectric arrangement. With respect to low-energy excitations the system may be treated as a 2D system of Heisenberg spins in an external magnetic field (bias),⁴⁵ known to describe the formation of ferromagnon-like quasiparticles that can propagate through the array.

At zero applied bias the fluctuations of electron density can cause spontaneous dipole formation. Let us set the probability of a dipole formation at the site i as χ_α^i , where the index $\alpha = 1, 2$ corresponds to different dipole orientations. The increment in the energy of the array due to formation of a dipole is

$$\Delta E = \sum_{i,\alpha} E_i \chi_\alpha^i - \frac{1}{2} \sum_{i,\alpha} \sum_{j,\beta} J_{ij}^{\alpha\beta} \chi_\alpha^i \chi_\beta^j$$

where $J_{ij}^{\alpha\beta}$ is the interaction energy. The energy of an individual dipole E_i depends on the scale of charge fluctuations q . In the mean field approximation $x_1 = \langle \chi_1^i \rangle$, $x_2 = \langle \chi_2^i \rangle$ and the energy increment per site is reduced to⁴⁶

$$\Delta E' = \bar{E}(x_1 + x_2) - \frac{1}{2} V(x_1^2 + x_2^2) + W x_1 x_2$$

where $V = V(n)$ corresponds to a self-action, $W = W(n)$ is an effective interaction between the dipoles of different orientation, and $n = q/e$ is the measure of charge fluctuation. The change of configurational entropy is found to be

$$\Delta S = \log \left[\frac{\sigma^{N(x_1+x_2)} N!}{(x_1 N)! (x_2 N)! (N(1-x_1-x_2))!} \right]$$

where N is the number of wires in the array and $\sigma = (2n)!/(n!)^2$ accounts for the degeneracy of a particular charge state. The equilibrium values of x_1 and x_2 are associated with a minimum in the increment of free energy, which yields

$$2\sigma \frac{1-\xi}{\sqrt{\xi^2 - \eta^2}} = \exp(\alpha - \gamma\xi/2), \quad \frac{\xi + \eta}{\xi - \eta} = \exp(\beta\eta)$$

where $\xi = x_1 + x_2$ is the concentration of dipoles, $\eta = x_1 - x_2$ is the total polarization, $\alpha = \bar{E}/T$, $\beta = (W + V)/T$, and $\gamma = (W - V)/T$ ($k_B = 1$). From the above one obtains:

$$\xi = \frac{1}{\beta} \frac{\xi + 1}{\xi - 1} \log(\xi)$$

where

$$\zeta = -1 + \frac{1}{2} \delta^2 \left[1 - \sqrt{1 - (2/\delta)^2} \right], \quad \delta = \frac{1}{\sigma} \frac{\xi}{1 - \xi} \exp(\alpha - \gamma \xi / 2)$$

The graphical solution of the equation defining ξ is shown in Fig. 3(a) for several different values of $W = V = \bar{E}/2$. The transition to the polarized state with change in interaction parameter values is manifested as a jump in the equilibrium value of dipole concentration ξ from zero to $\xi \sim 0.8$.

In the dipole model of close-packed nanowire arrays, we see many analogies with cellular automata: the state of a given cell (or dipole, or nanowire) is determined partly by global external forces — an external voltage bias, and partly by the local environment — the state of adjacent dipoles. Such interactions in the context of cellular automata and also neural networks are the subject of much research effort in computer science. While current computer systems implement these models entirely in software, with some effort made on the architecture level, a computer based on the physics of interacting dipoles would solve such problems directly through massively parallel physical interactions, and much more rapidly than software running on conventional computer systems.

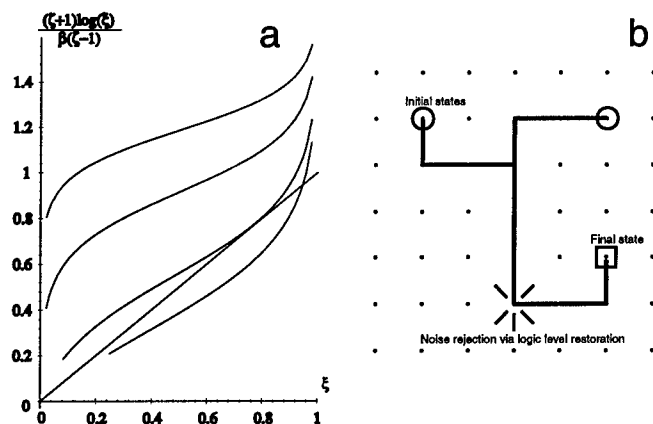


Figure 3. (a) Graphical solution of governing equation for ξ , showing the transition to a polarized state indicated by a jump in equilibrium ξ . The parameter values are $n = 7$, $W = V = \bar{E}/2 = 20, 15, 11.3$, and 10 . (b) State representation of a digital computer showing binary level restoration (noise rejection). The analog computer operates in the same manner in a continuous state space.

4. Computation schemes employing ordered nano-arrays

Though the sequential binary processing paradigm used in existing computers was formulated by von Neumann, it is interesting to note that in some of his later publications, he took a radically different direction from his previous work.⁴⁷ He had grown interested in a form of massively-parallel distributed computation, inspired by the remarkable computations performed by the human brain. This computational architecture is premised on the idea that a variety of sophisticated computations may be performed by a large coupled array of simple computational elements obeying basic rules of behavior — an ideal opportunity for the densely packed, ordered nano-arrays that we are now able to fabricate. The computational potential of such systems for particular classes of problems was recognized early on by Hopfield.⁴⁸

- *Potential advantages of physical massively parallel computation*

Distributed computational systems are known for their ability to solve rapidly difficult computational problems such as pattern recognition or path optimization (e.g. "travelling salesman") through massively parallel operation of (possibly slow) computational elements. Such systems are also fault-tolerant due to their "graceful" degradation with the failure of individual computational elements. Fault-tolerance is an important consideration in view of the imperfect statistical yield of nanoscale components and systems produced by both conventional lithography and by non-lithographic means (Section 1). Thus, massively parallel, neural computing is the approach to computation naturally suited for building computers from self-organized, non-lithographic nanomaterial systems.

- *Computation as evolution of phase space trajectory*

In seeking novel computational capabilities derivable from ordered nanoarrays, it is useful to examine a generalized picture of computational machines. We describe a computational machine by a space of states and system dynamics acting in the space. Each state vector contains all relevant information about the system, including the state of all computational elements, memory elements, inputs, and outputs. The system dynamics describe the evolution of the computer from one state to the next. Computations are therefore represented as trajectories in a state space or phase space. We restrict ourselves to systems where the dynamics are deterministic, treating noise as a stochastic perturbation to the state of the system.

For example, in this framework, the digital computer is a machine having a discrete state space and discrete dynamics, iteratively mapping the state f_i to the state f_{i+1} until a stable point (solution) is reached — see Fig. 3(b). Computation is performed by the deterministic transition from one state to the next, ultimately mapping every initial (input) state to a final (output) state. The discreteness of the state space arises from logic level restoration, a fundamental requirement of digital computation, which restores state information to the nearest grid point.

In contrast, the distributed computer possesses a continuous state space (with dimension equal to the number of computational elements) and a differential dynamics over that space. Again, computation is embodied in the evolution of the system state: i.e. in the mapping formed between initial and final states. For this computational machine to be capable of rejecting noise—the stochastic perturbation of the system state to some neighboring state—adjacent initial states $f(x)$ and $f(x + \delta)$ must result in identical outputs. In other words, noise rejection demands differential stability for all state space trajectories. Conveniently, this constraint and its implications have already been studied in detail in the work of Hopfield, thus providing a foundation for the study of computational systems of the type examined here.

- *First order dynamics of computational machines*

Although the physical system implementing the desired computational machine may be quite complex, it is often sufficient, and certainly illustrative, to model it as an array of interacting first-order elements. In the language of neural networks, we define S_i as the state of element i , as determined by the value of some observable quantity, and q_i as its activation potential, representing an internal parameter more directly related to the inputs to i . In general, the inputs to an element represent the collective influence of other computational elements in the array combined with a possible external force. The state S_i and the activation potential are related by $S_i = \Phi_i(q_i)$ where the activation function Φ_i is often required to be monotonic, but need not be identical for all computational elements. The continuous-time deterministic system is then guided by the dynamics given by

$$C_i \frac{dq_i}{dt} = \sum_j w_{ij} s_j - R_i q_i - \theta_i$$

where C_i represents the response time of the element (i.e. capacitance), R_i is a gain parameter, θ_i is the external input to the element, and w_{ij} is the weight for coupling element i to element j . The steady state solution of this equation reveals that the activation potential of an element is determined by a weighted linear superposition of the internal and external inputs.

The stability condition necessary for noise rejection in the distributed processor takes a simple form in this context, requiring that there be symmetric coupling between the elements, i.e. $w_{ij} = w_{ji}$. Furthermore, this coupling symmetry, which is fortunately the rule rather than the exception in physical systems, dictates that the computational trajectories are the gradients of the energy function given by⁴⁹

$$E = -\sum_{i \neq j} w_{ji} s_i s_j + \sum_{j=1}^N \frac{1}{R_j} \int_0^{s_j} \varphi_j^{-1}(s_j) ds_j + \sum_{j=1}^N \theta_j s_j$$

Similar results can be obtained for more general systems, including those for which the dynamics contain a stochastic (thermal) character. Such an energy function is reminiscent of the array of dipoles as discussed in Section 3. Whether such phenomena can endow controllable nano-array systems with nontrivial

computational functions is still an open question. This issue has been addressed by neural network researchers, albeit from the opposite direction of seeking a *model* system in which to implement a computer and then creating (discovering) the symmetrically coupled array through the process of training.

- *Neural networks and self-organized nanostructure arrays*

Pursuing further the analogy with neural network theory, we inherit a roadmap for applications, approaches, and potential pitfalls. From work on the Hopfield machine came applications in error correction, associative memory, energy minimization, and an awareness of the problems presented by spurious local minima in the energy surface. In the Boltzmann and mean-field-theory machines we find an example of traditional input-output based computing, as well as a guide to non-zero-temperature computation.⁵⁰

The analogy with neural networks is not a perfect one. Considerable work in neural network theory is concerned with network training — that is, adjusting the weights w_{ij} of the connections based on the difference between two solutions to a representative problem: that obtained by the network, and the correct solution. The practical difficulties associated with training have seriously limited the popularity and applicability of this computational approach. In the case of ordered nano-arrays, however, the inter-element weights or couplings may be fixed and the system becomes untrainable in the conventional sense. Still, a non-adaptive machine based on such a system can still offer many advantages including low power consumption, fault tolerance, and possibly wireless implementation; the set of problems solvable by such a system is, of course, reduced.

Although neural network theory usually favors adjustment of the weights w_{ij} , neglecting the possibility of adjusting individual (nonlinear) activation functions, $\Phi_i(q_i)$, the latter may be more attractive technologically. For example, selectively loading nanopores with different materials could alter the conductivity of the nanowires, or the response of their polarization to external influences; a lithographic writing step could produce contacts to separately bias individual nanowires or small regions of the array. In general, some activation functions could be temporarily or permanently differentiated from others. These are some of the possible avenues for engineering complexity into the array. With increasing control over the fabrication of individual nanostructures, more general and powerful functions will be possible.

Uniformity and/or symmetry in the array is both a limitation and a potential strength — while lattice translation symmetry indicates low information content, it also represents an opportunity for the creation of powerful and flexible systems — if the regions of the array may be dynamically differentiated from others, it may be possible to create/program a computer "on the fly" that will solve a given computational problem, read the result, and then generate immediately a different computer for the next problem. Such speculations presuppose a very high degree of control over the writing and reading of different states to and from the individual elements of the array, and obviously demand extremely fast, high bandwidth communication capability to simultaneously program arbitrary states

into all elements of the array. While highly speculative at the present time, these objectives could be seen as the ultimate long-term goals for the nanocomputational architect or system designer — to step out of the way of the physics as much as possible, and let computation proceed via fast and fundamental processes.

5. Conclusions

Miniaturization is the primary source of the remarkable applications enabled and created by microelectronics. However, because microelectronics still operates according to VLSI design rules premised on macroscopic physics, it is ill equipped to deal with the sub-100 nm realm. Even if the technical challenges can be overcome, the economics of Moore's law are becoming increasingly unattractive. Since the SIA roadmap predicts minimum feature sizes on the 100 nm scale within the next ten years, it is important to study potential successor technologies. These technologies, instead of merely dealing with the problems associated VLSI miniaturization, can instead be designed from the ground up to embrace and exploit the fundamental physics of electrons and atoms on the nanoscale. Based on their low cost, high manufacturing throughput, and material versatility, it is non-lithographically fabricated nanosystems that will spawn such technologies, which will require new architectures incorporating fault tolerance, high interconnection, and possibly a neural network approach rather than serial binary logic. However, the cooperative behavior and collective phenomena of self-organized nanodevice arrays are shown to be compatible with these basic requirements. While the challenges facing the implementation of such technologies are significant, we must remember that no one (except for Gordon Moore) predicted the pace of evolution in planar integration even after several years of development. Who can say now what the future holds for non-lithographic nanosystems, or what applications they might enable?

References

1. B. Hoeneisen and C. Mead, "Fundamental limitations in microelectronics. I. MOS technology," *Solid State Electron.* **15**, 819 (1972);
R. H. Dennard, F. H. Gaensslen, H. N. Yu *et al.*, "Design of ion-implanted MOSFETs with very small physical dimensions," *J. Solid State Circ.* **9**, 256 (1974).
2. Semiconductor Industry Association, *The National Technology Roadmap for Semiconductors*, 1997.
3. M. S. Malone, *The Microprocessor: A Biography*, New York: Springer-Verlag, 1995.

4. See, for example, H. Kawaura, T. Sakamoto, T. Baba, *et al.*, "Transistor operation of 30-nm gate-length EJ-MOSFETs," *IEEE Electron Dev. Lett.* **19**, 74 (1998) and references therein.
5. L. Geppert, "The media event: Moore's Law mania," *IEEE Spectrum* **35**, 20 (1998).
6. R. W. Keyes, "The future of the transistor," *Sci. Amer. Special Issue* **8**, 48 (December, 1997).
7. S. Haykin, *Neural Networks: A Comprehensive Foundation*, New York: IEEE Press, 1994.
8. C. S. Lent and P. D. Tougaw, "Device architecture for computing with quantum dots," *Proc. IEEE* **85**, 541 (1997), and references therein.
9. S. Datta, *Electronic Transport in Mesoscopic Systems*, Cambridge, U.K.: Cambridge University Press, 1995;
D. K. Ferry and S. M. Goodnick, *Transport in Nanostructures*, Cambridge, U.K.: Cambridge University Press, 1997.
10. J. R. Heath, P. J. Kuekes, G. S. Snider, and R. S. Williams, "A defect-tolerant computer architecture: opportunities for nanotechnology," *Science* **280**, 1716 (1998).
11. F. Burmeister, C. Schaffe, B. Keilhafer, *et al.*, "From mesoscopic to nanoscopic surface structures: lithography with colloid nanolayers," *Adv. Mater.* **10**, 495 (1998).
12. S. Y. Chou, P. R. Krauss, and L. Long, "Nanolithographically defined magnetic structures and quantum magnetic disk," *J. Appl. Phys.* **79**, 6101 (1996).
13. R. Notzel, "Self-organized growth of quantum-dot structures," *Semicond. Sci. Technol.* **11**, 1365 (1996).
14. D. Levy and L. Esquivias, "Sol-gel processing of optical and electrooptical materials," *Adv. Mater.* **7**, 120 (1995).
15. P. P. Nguyen, D. Pearson, and R. Tonucci, "Fabrication and characterization of uniform metallic nanostructures using nanochannel gas," *J. Electrochem. Soc.* **145**, 247 (1998).
16. V. M. Cepak, J. C. Hulteen, G. Che, *et al.*, "Chemical strategies for template synthesis of composite micro- and nanostructures," *Chem. Mater.* **9**, 1065 (1997).
17. D. Routkevitch, A. A. Tager, J. Haruyam, *et al.*, "Nonlithographic nano-wire arrays: fabrication, physics, and device applications," *IEEE Trans. Electron Dev.* **43**, 1646 (1996).
18. F. Keller, M. S. Hunter, and D. L. Robinson, "Structural features of oxide coatings on aluminum," *J. Electrochem. Soc.* **100**, 411 (1953).
19. J. P. O'Sullivan and G. C. Wood, "The morphology and mechanism of formation of porous anodic films on aluminum," *Proc. Roy. Soc. Lond. A* **317**, 511 (1970).

20. H. Masuda and K. Fukuda, "Ordered metal nanohole arrays made by a two-step replication of honeycomb structures of anodic alumina," *Science* **268**, 1466 (1995);
H. Masuda, F. Hasegawa, S. Ono, "Self-ordering of cell arrangement of anodic porous alumina formed in sulfuric acid solution," *J. Electrochem. Soc.* **144**, L127 (1997).
21. O. Jessensky, F. Muller, and U. Gosele, "Self-organized formation of hexagonal pore arrays in anodic alumina," *Appl. Phys. Lett.* **72**, 1173 (1998).
22. L. Zhang, H. S. Cho., F. Li, R. M. Metzger, and W. D Doyle, "Cellular growth of highly ordered porous anodic films of aluminium," *J. Mat. Sci. Lett.* **17**, 291 (1998).
23. D. AlMawlawi, N. Coombs, and M. Moskovits, "Magnetic properties of Fe deposited into anodic aluminum oxide pores as a function of particle size," *J. Appl. Phys.* **70**, 4421 (1991).
24. G. Hornyak, M. Kroll, R. Pugin, *et al.*, "Gold clusters and colloids in alumina nanotubes," *Chem-Eur. J.* **3**, 1951 (1997).
25. Z. Zhang, J. Y. Ying, and M. Dresselhaus, "Bismuth quantum-wire arrays fabricated by a vacuum melting and pressure injection process," *J. Mater. Res.* **13**, 1746 (1998).
26. J. Li, M. Moskovits, and T. L. Haslett, "Nanoscale electroless metal deposition in aligned carbon nanotubes," *Chem. Mater.* **10**, 1963 (1998).
27. R. V. Parthasarathy, K. L. N. Phani, and C. R. Martin, "Template synthesis of graphitic nanotubules," *Adv. Mater.* **7**, 896 (1995).
28. A. A. Tager, D. Routkevitch, J. Haruyama *et al.*, "Nonlithographic fabrication and physics of nanowire and nanodot array devices – present and future," in: S. Luryi, J. M. Xu and A. Zaslavsky, eds., *Future Trends in Microelectronics*, Dordrecht: Kluwer, 1996.
29. G. A. Ozin, "Nanotechnology: synthesis in diminishing dimensions," *Adv. Mater.* **4**, 612 (1992).
30. R. J. Tonucci, B. L. Justus, A. J. Campillo, and C. E. Ford, "Nanochannel array glass," *Science* **258**, 783 (1992).
31. R. Leon, P. M. Petroff, D. Leonard, and S. Fafard, "Spatially resolved visible luminescence of self-assembled semiconductor quantum dots," *Science* **267**, 1666 (1995).
32. T. A. Jung, R. R. Schlittler, J. K. Gimzewski, H. Tang and C. Joachim, "Controlled room-temperature positioning of individual molecules: molecular flexure and motion," *Science* **271**, 181 (1996).
33. R. E. Scholten, J. J. McClelland, E. C. Palm, A. Gavrin and R. J. Celotta, "Nanostructure fabrication via direct writing with atoms focused in laser fields," *J. Vac. Sci Technol. B* **12**, 1847 (1994).
34. G. M. Whitesides, "Self-assembling materials," *Sci. Amer.* **273**, 146 (1995).

35. S. W. Pang, T. Tamamura, M. Nakao, A. Ozawa, and H. Masuda, "Direct nano-printing on Al substrate using a SiC mold," *J. Vac. Sci Technol. B* **16**, 1145 (1998).
36. P. M. Ajayan and T. W. Ebbesen, "Nanometre-size tubes of carbon," *Rep. Prog. Phys.* **60**, 1025 (1997).
37. C. T. White, D. H. Robertson, and J. W. Mintmire, "Helical and rotational symmetries of nanoscale graphitic tubules," *Phys. Rev. B* **47**, 5485 (1993).
38. J. W. G. Wildoer, L. C. Venema, A. G. Rinzler, R. E. Smalley, and C. Dekker, "Electronic structure of atomically resolved carbon nanotubes," *Nature* **391**, 59 (1998).
39. A. Yu. Kasumov, H. Bouchiat, B. Reulet *et. al.*, "Conductivity and atomic structure of isolated multiwalled carbon nanotubes," *Europhys. Lett.* **43**, 89 (1998).
40. R. Landauer, "Self-polarized quantum dots," *Solid State Commun.* **95**, 7 (1995).
41. K. Kempa, D. A. Broido, and P. Bakshi, "Spontaneous polarization in quantum-dot systems," *Phys. Rev. B* **43**, 9343 (1991).
42. K. Kempa, P. Bakshi, D. A. Broido, *et al.*, "Polarization of electrons in quantum dashes in magnetic field," *Solid State Commun.* **91**, 231 (1994).
43. M. Garzon, *Models of Massive Parallelism: Analysis of Cellular Automata and Neural Networks*, Heidelberg: Springer-Verlag, 1995);
S. Wolfram, *Cellular Automata and Complexity: Collected Papers*, New York: Addison-Wesley, 1994.
44. V. A. Sablikov and B. S. Shchamkhalova, "Coulomb interaction during coherent transport of electrons in quantum wires," *JETP Lett.* **67**, 196 (1998).
45. J. M. Ziman, *Principles of the Theory of Solids*, Cambridge, U.K.: Cambridge University Press, 1972.
46. M. E. Lines and A. M. Glass, *Principles and Applications of Ferroelectric and Related Materials*, Oxford: Clarendon Press, 1977.
47. J. von Neumann, *Theory of Self-Reproducing Automata*, Chicago: Univ. of Illinois Press, 1966.
48. J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proc. Nat. Acad. Sci.* **79**, 2554 (1982).
49. J. J. Hopfield, "Neurons with graded response have collective computational properties like those of two-state neurons," *Proc. Nat. Acad. Sci.* **81**, 3088 (1984).
50. G. E. Hinton and T. J. Sejnowski, in *Parallel Distributed Processing: Explorations in Microstructure of Cognition*, S. J. Hanson J. D. Cowan and C. L. Giles, eds. (MIT Press: Cambridge, 1986).

Molecular-Scale Electronics

M. A. Reed

Dept. of Electrical Engineering, Yale University, New Haven, CT 06520 USA

1. Introduction

It is well recognized that conventional VLSI lithography-based technology is fast approaching the limits of its capabilities. The underlying issues responsible are numerous, but fundamentally result in increasingly difficult and expensive lithography.¹ To surmount these problems, nanoscale quantum devices and circuits have been proposed.^{2,3} Over the last two decades, demonstrations of many of these technologies have been accomplished.⁴ These demonstrations include RTD and RTT devices and circuits that promise compact multi-valued logic and memories; quantum dot and single electron devices; and others. Will these be viable technology alternatives for the post-VLSI era?

Present embodiments of these technologies demand tremendous control over the $I(V)$ characteristics of the device, an undesirable property for downscaling. Dimensional control is a dominant obstacle, since nanodevices must operate by tunneling in some fashion. Since a barrier is needed for isolation in a 3-terminal device with gain, tunneling will be exponentially sensitive to atomic-layer fluctuations in barriers, resulting in device-specific variations that may be unacceptable for large-scale integration. A related problem is that devices using discrete electron charging (SETs) only work at reduced temperatures. Robust room temperature operation (leakage requirements similar to VLSI requirements) requires junctions smaller than 1 nm, which will thus suffer severe tunneling fluctuations. Last but not least is that these embodiments do not critically address the major limiting factors of a 2D lithography-based technology: accessible parallel fabrication, interconnection density, and alignment. Any successful new technology must (1) solve the interconnect problem, (2) use self-aligned fabrication, and (3) operate at room temperature and at the atomic level. Scaling any technology to the 10 nm level may not be cost-effective, as the performance increase is marginal compared to development costs. An atomic- or molecular-scale technology may be the only approach worth the investment.

Recently, molecular electronics-based computation has attracted attention, because it addresses the ultimate in a dimensionally scaled system; ultra-dense and molecular-scale.^{5,6} The significant scaling factor gained from molecular-scale devices implies eye-opening comparisons: a contemporary computer utilizes $\sim 10^{10}$ silicon-based devices, whereas one could prepare $\sim 10^{23}$ devices in a single beaker. An additional driving factor is the potential to utilize directed self-assembly of components⁷ such as chemically synthesized interconnects, active devices, and circuits. This novel technological approach for post-VLSI electronic systems can

conceivably lead to a new era in ultra-dense electronic systems. Spontaneous assembly of atomic-scale electronics attacks the interconnection and critical dimension control problems in one step. Concurrently, the approach utilizes self-aligned batch processing techniques that address the fabrication limitations of conventional ULSI.

Molecular (i.e., organic) materials for electronic and optoelectronic applications have been realized for quite some time. In addition to uses such as liquid crystal displays and organic photoresists, devices such as light-emitting diodes and lasers, transistors, and sensors have been demonstrated.⁵ The distinction between these (essentially "bulk") applications and molecular-scale electronics is not just one of size, but of concept: the design of a molecule that itself is the active element.

Molecules were proposed as active electronic devices as early as 1973, when Aviram and Ratner⁸ proposed that unimolecular rectification, or asymmetrical electrical conduction, should occur through the molecular orbitals of a single D- σ -A molecule by "through-bond tunneling". Here D is an electron donor with low ionization potential, A is an electron acceptor with high electron affinity, and σ is a covalent "bridge". The excited zwitterionic state D⁺- σ -A⁻ would be relatively accessible from the ground neutral state D- σ -A, while the opposite zwitterion D⁻- σ -A⁺ would lie several eV higher and be inaccessible. In solid-state language, the system is an asymmetric resonant tunneling structure. Indeed, molecular systems have good analogies to solid state systems. Instead of the Fermi levels of the solid state, one deals with the highest occupied and lowest unoccupied molecular orbitals (HOMO and LUMO, respectively) of molecules. Instead of metal and interconnects, one uses conjugated linear polymeric systems. Instead of "doping" to modify Fermi levels, one modifies the electron affinity and ionization potential of molecules by the chemical substitution. And by designing in the molecular orbitals, one has the equivalent of bandgap engineering.⁴

For molecular-scale electronics to come of age, fabrication and measurement techniques had to reach the atomic scale. The advent of atomic imaging techniques, such as the scanning tunneling microscope (STM) and the atomic force microscope (AFM), have given us an atomic view of molecular placement and self-assembly. Nanofabrication techniques have created interconnects small enough to reliably contact molecules. An understanding of electronic transport at the atomic level has developed over the last decade, and theoretical models of conduction through such systems are beginning to develop. All these advances have led to the first electrical measurements of molecular systems. Among these are conductivity measurements of molecules by STM,⁹ and the first measurements of electronic conduction through a single "molecular wire"¹⁰ and molecule.^{11,12} This paper reviews some of the initial work in this new, exciting field.

2. Self-assembled conjugated molecules

It is well-documented that bulk conjugated organic materials can be semiconducting, or conducting, when doped.¹³ However, the measurements of

long distance electron transfer through single molecules are now just emerging. Candidates for molecular wires and switches include polyporphyrins, polyphenylenes, polythiophenes, and other planar organic polymers with extended π -conjugation (i.e. electron delocalization along the length of the molecule, which can be verified by optical measurements). A review of the synthesis of conjugated oligomers can be found elsewhere,^{5,15} as well as a general review of candidate molecular conductors.^{6,16-18} Examples of some of these oligomers are shown in Fig. 1.

Although there is a variety of synthetic methods with their own various advantages, one of the very attractive properties of the oligomers shown in Figure 1 is a high degree of purity (i.e., of dimensional control). In many synthetic processes, separation of an n -mer from an $(n + 1)$ -mer limits purity. An alternate recent approach (producing the structures of Fig. 1) is to synthesize using a divergent-convergent method;^{19,20} thus the separation becomes that of an n -mer versus a $2n$ -mer, a vastly more efficient and rapid process that produces monodisperse, stable, and soluble conjugated oligomers.

The challenge that has stymied the entire field for more than 2 decades has been the elementary task of "soldering" one of these electrically active candidate

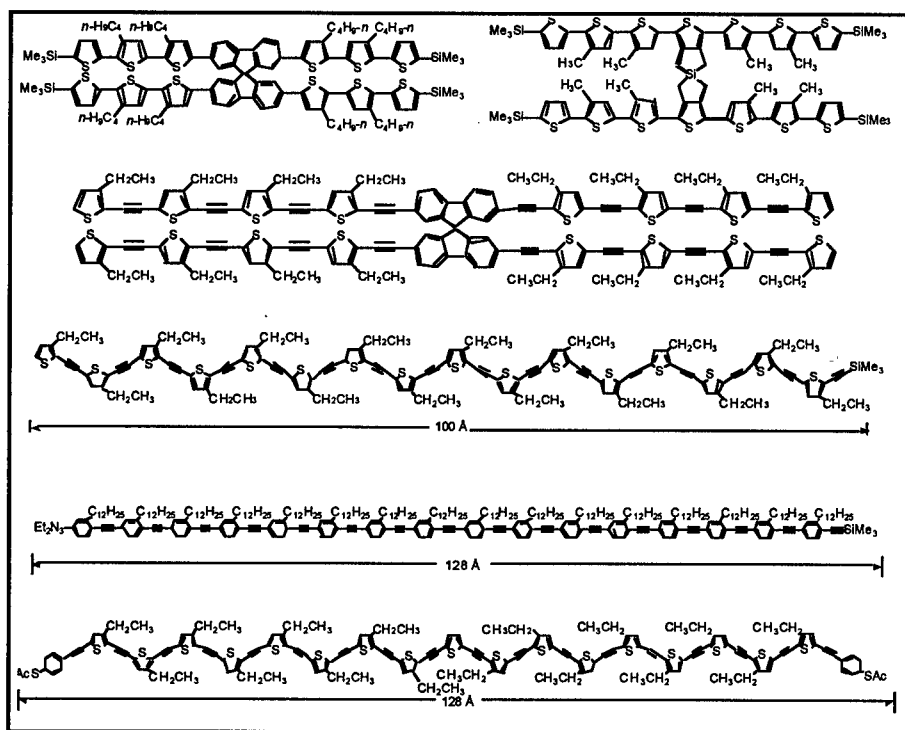


Figure 1. A variety of molecular conductors. The top three are conductive oligomers connected to an orthogonally-fused crossbar; the bottom two are long rigid-rod conjugated oligomers. Note the thiol termini on the bottom structure.

molecules between contacts. With the development of atomic imaging, one can now imagine using STM or AFM tips to maneuver, by brute force, the molecules into the correct position. A much more attractive alternative is the spontaneous self-absorption of the molecular species between the contacts.²¹ Over the last decade, the ability to form very well defined self-assembled monolayers (SAMs) of oligomers on metal has been demonstrated. The most widely studied system has been the Au-SR system ($R = \text{alkyl}$), which forms very well ordered single monolayers with a very aggressive ~ 2 eV bond. In fact, the bonding strength is so large that there is evidence the SAM actually "anneals out" defects.

Functional terminal moieties such as thiols now provide the key to self-assembled structures. By synthesizing these onto the oligomer end-groups^{22,23} such as the bottom oligomer example in Fig. 1, we have a self-absorbing and orienting species (the thioacetyl end groups ($\text{SAc} = \text{SCOCH}_3$) serve as a protected self-assembled chemisorbed termini with *in-situ* generation of the free thiol). This species allows us to investigate conduction through conjugated molecules that are end-bound onto a surface, or placed between proximal probes (e.g., for gold probes one uses thiol end-groups). In the last few years, a variety of metal (or semiconductor)surface/functional end-group combinations have been investigated, giving design flexibility in both synthetic design and multi-terminal interconnect orientation.⁷ This technique has yielded the first electrical measurements of molecular systems.

3. Conductivity measurements of molecular systems

The measurement of charge transport in single organic molecules, and the determination of their conductance, has been a long sought goal. Such measurements are experimentally challenging and intriguing since one can test the validity of transport approximations at the molecular level. A conceptually simple configuration would be to connect a single molecule between metallic contacts. Such a metal-molecule-metal configuration would present the molecular embodiment of a system analogous to a quantum dot,²⁴⁻²⁷ with the potential barriers of the semiconductor system replaced by the contact barrier of the molecule/metal interface. Although making nanofabricated planar contacts at the molecular scale is challenging, it has been done and candidate molecular conductors have been absorbed into the gaps.²⁸ An absence of observable conductivity in these systems may be due to inefficient electron transfer, undesirably large contact potentials, conjugation-breaking due to substrate interactions, or unfavorable absorption configurations. Thus, a number of alternative approaches have been attempted and realized.

The first obvious approach is to use STM techniques to locate and locally measure candidate oligomers absorbed on a (metallic) surface. However, the oligomers will not preferentially be near surface-normal in isolation. By using an alkanethiol (i.e., inert non-conductive) SAM as an inert host matrix, single conjugated molecules have been inserted (randomly at defect sites) and imaged⁹ (Fig. 2). It was observed that the molecules had a significantly higher

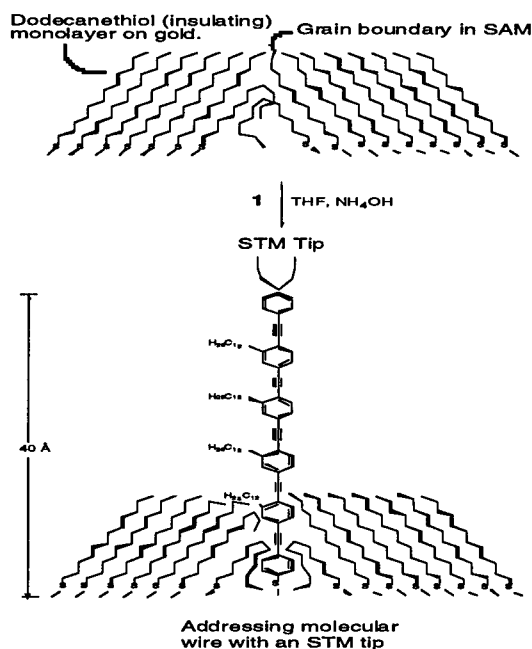


Figure 2. Schematic STM experimental set-up to observe the conductivity in π -conjugated molecules.

conductivity that the alkanethiol background; however these experiments could not give the absolute value of the conductivity due to the involvement of a tunneling gap between the STM tip and the molecule. One variant of this approach was to use Au nanoclusters on a doubly-functionalized oligomer^{29,30} and by STM measurement deduce the molecule conductivity (although the uncertainty in cluster size resulted in large error bars). Another variant has been to measure C_{60} as the intermediate species.³¹ We exclude C_{60} , carbon nanotubes, and other related structures as they do not have the desirable synthetic and functionalization properties that address the fabrication issues discussed previously. The ideal embodiment is to create statically stable contacts, while concurrently restricting the number of active molecules to as few as one. This goal has been realized¹⁰ (using benzene-1,4-dithiolate connected between stable proximal metallic gold contacts) using a mechanically controllable break junction (MCB)^{32,33} (Fig. 3). A notched metal wire is glued onto a flexible substrate; and is fractured by bending the substrate, after which an adjustable tunneling gap can be established. A large reduction factor between the piezo elongation and the electrode separation ensures an inherently stable contact or tunnel junction. The wire contacts are atomically sharp when broken, as demonstrated by conductance quantization. The deposition

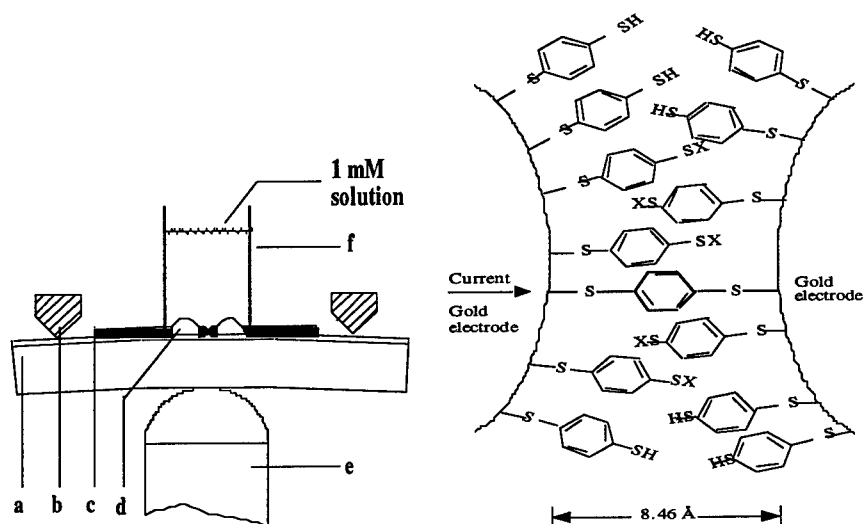


Figure 3. (left) A schematic of the MCB junction with (a) the bending beam, (b) the counter supports, (c) the notched gold wire, (d) the glue contacts, (e) the piezoelement, and (f) the glass tube containing the solution. (right) A schematic of a benzene-1,4-dithiolate SAM between proximal gold electrodes formed in a MCB. X can be either H or Au.

onto the contacts resulted in formation of a SAM on the gold electrodes nearly perpendicular to the surface.³⁴

Current-voltage $I(V)$ and conductance $G(V)$ measurements showed reproducible characteristic features of stepped $G(V)$ with a first step of 22.4 ± 0.3 M Ω . This figure is compared to a resistance of ~ 9 M Ω and 18 ± 12 M Ω deduced from the Au nanocluster experiments.^{29,30} An interpretation of the observed ~ 0.7 V gap is due to the mismatch between the contact Fermi level and the LUMO. Preliminary calculations using this interpretation give similar characteristics to the experimentally observed data.³⁵

Direct contacting to organic thin films has also been done. However, these experiments involve multilayers of molecular wires and micron-scale device areas containing a large number of molecules, which complicates the analysis of the transport mechanism of single molecules. Experiments with an evaporated-metal-top-contact/molecules/metallic-bottom-contact configuration, which has tens of thousands of parallel active molecules, have been demonstrated.³⁶ One experiment on an organic system³⁷ reported evidence for Coulomb charging. Direct contacting to and electrical measurements of a single monolayer of a small number of molecules are needed for accurate spectroscopy.

A novel fabrication technique has been reported to directly measure the conduction through a small number of organic molecules.¹² These devices, called

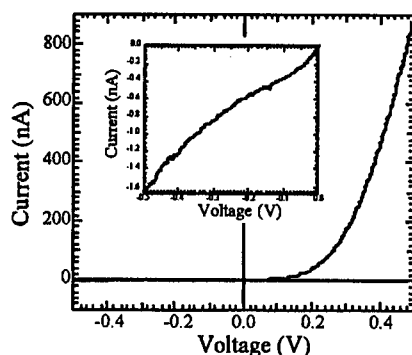
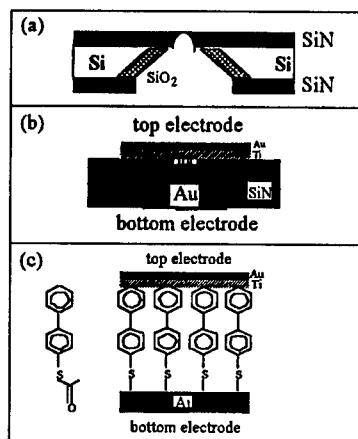


Figure 4. (left) Fabrication of nanopores. (a) Cross section of a silicon wafer showing a ~ 300 Å pore etched in a SiN membrane; (b) Au-Ti/SAM/Au structure, (c) specific SAM (4-thioacetylphenyl) and structure schematic. (right) Diode-like $I(V)$ characteristics at room temperature (negative bias polarity magnified in the inset).

"nanopores", consist of a self-assembled monolayer of conjugated molecular wires sandwiched between top and bottom metallic contacts. The employment of nanoscale device size insures that the adsorbed organic layer is highly ordered and defect-free. A schematic of the device is shown in Fig. 4. This technique gives a yield of about 80% working and stable devices.

Prominent rectifying behavior is observed: the current at 1 V bias is about 500 times higher than the current at -1 V bias. This behavior is distinctly different from the Aviram-Ratner unimolecular rectifier⁸ (rectification due to an asymmetric ground-state/excited state zwitterion), recently experimentally observed in hexadecylquinolinium tricyanoquinodimethanide Langmuir-Blodgett films.¹¹ In the present case, the asymmetry is simply produced by the different contact barriers, loosely analogous to Schottky barriers. While the $I(V)$ curve at negative bias is rather linear,³⁸ the $I(V)$ curve at positive bias displays exponential behavior, with $\ln(I) \propto 1/T$ and the slopes a function of the bias voltage. Using standard thermal emission theory, a value of 0.26 eV for the direct metal-molecule barrier is found. This technique lends itself to rapid synthetic and metallic variation, and is expected to be a workhorse in uncovering a wealth of metal-molecule transport characteristics previous inaccessible by other techniques.

4. Conclusion

Discussed here are the basic concepts of self-assembled molecular-scale electronic devices. New fabrication techniques to study the transport mechanism of organic molecular wires, such as MCBs and nanoscale fabrication with special metal

deposition techniques, are employed to provide metallic contacts to the self-assembled monolayer of a small number of conjugated molecular wires. In one case we have measured the conductivity of a single molecule. These methods can be easily adapted to other molecular wire systems for determination of the transport mechanism and band alignment. A clear and outstanding challenge is the synthesis and measurement of three-terminal structures, and the determination of gain mechanisms in such structures.

5. Acknowledgments

I am indebted to the dedicated work of my students and postdocs C. Zhou, C. J. Muller, D. L. Lombardi, J. Chen, and fabrication support from J. Sleight and M. Deshpande. I am very grateful for the stimulating collaborations with J. M. Tour and D. L. Allara, and their respective research groups; and stimulating discussions with M. Ratner. I gratefully acknowledge DARPA for financial support under ONR grant N00014-95-1-1182.

References

1. R. W. Keyes, "The future of the transistor," *Science* **230**, 138 (1985); *Sci. Amer.* June 1993, pp. 70-78.
2. W. P. Kirk and M. A. Reed, eds., *Nanostructures and Mesoscopic Systems*, San Diego: Academic Press, 1992.
3. N. G. Einspruch and W. R. Frensley, eds., *Heterostructure and Quantum Devices*, San Diego: Academic Press, 1994.
4. C. Weisbuch and B. Vinter, *Quantum Semiconductor Structures*, San Diego: Academic Press, 1991.
5. J. Jortner and M. Ratner, eds., *Molecular Electronics*, Oxford, U.K.: Blackwell Science, 1997.
6. A. Aviram and M. Ratner, eds., *Molecular Electronics: Science and Technology*, New York: Annals of the New York Academy of Sciences, 1998.
7. D. L. Allara, T. D. Dunbar, P. S. Weiss, *et al.*, in: A. Aviram and M. Ratner, eds., *Molecular Electronics: Science and Technology*, New York: Annals of the New York Academy of Sciences, 1998, p. 349.
8. A. Aviram and M. A. Ratner, "Molecular rectifiers," *Chem. Phys. Lett.* **29**, 277 (1974).
9. L. A. Bumm, J. J. Arnold, M. T. Cygan, *et al.*, "Are single molecular wires conducting?" *Science* **271**, 1705 (1996).
10. M. A. Reed, C. Zhou, C. J. Muller, T. P. Burgin, and J. M. Tour, "Conductance of a molecular junction," *Science* **278**, 252 (1997).
11. R. M. Metzger, B. Chen, U. Höpfner, *et al.*, "Unimolecular electrical rectification in hexadecyl-quinolinium tricyanoquinodimethanide," *J. Am. Chem. Soc.* **119**, 10455 (1997).

12. C. Zhou, M. R. Deshpande, M. A. Reed, L. Jones II, and J. M. Tour, "Nanoscale metal/self-assembled monolayer/metal heterostructures," *Appl. Phys. Lett.* **71**, 611 (1997).
13. T. A. Skotheim, ed., *Handbook of Conducting Polymers*, New York: Dekker, 1986.
14. A. Aviram, ed., *Molecular Electronics: Science and Technology*, New York: American Institute of Physics, 1992.
15. D. L. Pearson, L. Jones II, J. S. Schumm, and J. M. Tour, "Molecular scale electronics. Syntheses and testing," *Synth. Metals* **84**, 303 (1997).
16. For some theoretical considerations on molecular-scale wires, see C. Joachim and J. F. Vinuesa, "Length dependence of the electrical transperence (conductance) of a molecular wire," *Europhys. Lett.* **33**, 635 (1996).
17. For some recent background work on the formation of molecular-based transporters and devices, see M. D. Ward, "Current developments in molecular wires," *Chem. Ind.* **1996**, 569.
18. M. C. Petty, M. R. Bryce, and D. Bloor, eds., *Introduction to Molecular Electronics*, New York: Oxford University Press, 1995.
19. J. S. Schumm, D. L. Pearson, and J. M. Tour, "Iterative divergent/convergent doubling approach to linear conjugated oligomers. A rapid route to a 128 Å long potential molecular wire," *Angew. Chem. Int. Ed. Engl.* **33**, 1360 (1994).
20. D. L. Pearson, J. S. Schumm, and J. M. Tour, "Iterative divergent/convergent approach to conjugated oligomers by a doubling of molecular length at each iteration. A rapid route to potential molecular wires", *Macromolecules* **27**, 2348 (1994).
21. P. E. Laibinis, G. M. Whitesides, D. L. Allara, *et al.*, "A comparison of the structures and wetting properties of self-assembled monolayers of n-alkanethiols on the coinage metal surfaces Cu, Ag and Au," *J. Am. Chem. Soc.* **113**, 7152 (1991);
C. D. Bain, J. Evall, and G. M. Whitesides, "Formation of monolayers by the coadsorption of thiols on gold: variation in the head group, tail group, and solvent," *J. Am. Chem. Soc.* **111**, 7155 (1989);
M. J. Robertson and R. J. Angelici, "Adsorption of aryl and alkyl isocyanides on powdered gold," *Langmuir* **10**, 1488 (1994);
J. I. Henderson, S. Feng, G. M. Ferrence, T. Bein, and C. P. Kubiak, "Self-assembled monolayers of dithiols, diisocyanides, and isocyanothiols on gold: 'chemically sticky' surfaces for covalent attachment of metal clusters and studies of interfacial electron transfer," *Inorg. Chim. Acta* **242**, 115 (1996);
J. J. Hickman, C. Zou, D. Offer, *et al.*, "Combining spontaneous molecular assembly with microfabrication to pattern surfaces: selective binding of isonitriles to platinum microwires and characterization by electrochemistry and surface spectroscopy," *J. Am. Chem. Soc.* **111**, 7271 (1989);
J. M. Tour, L. Jones II, D. L. Pearson, *et al.*, "Self-assembled monolayers and multilayers of conjugated thiols, α,ω -dithiols, and thioacetyl-containing adsorbates. Understanding attachments between potential molecular wires and gold surfaces", *J. Am. Chem. Soc.* **117**, 9529 (1995).

22. L. Jones II, J. S. Schumm, and J. M. Tour, "Rapid solution and solid phase syntheses of oligo(1,4-phenylene-ethynylene)s with thioester termini: molecular scale wires with alligator clips. Derivation of iterative reaction efficiencies on a polymer support", *J. Org. Chem.* **62**, 1388 (1997).
23. D. L. Pearson and J. M. Tour, "Rapid syntheses of oligo(2,5-thiophene-ethynylene)s with thioester termini: potential molecular scale wires with alligator clips", *J. Org. Chem.* **62**, 1376 (1997)..
24. M. A. Reed, J. N. Randall, R. J. Aggarwal, *et al.*, "Observation of discrete electronic states in a zero-dimensional semiconductor nanostructures," *Phys. Rev. Lett.* **60**, 535 (1988).
25. L. P. Kouwenhoven, N. C. Van-der-Vaart, A. T. Johnson, *et al.*, "Single electron charging effects in semiconductor quantum dots," *Z. Phys. B* **85**, 367 (1991).
26. M. Tewordt, L. Martin-Moreno, V. J. Law, *et al.*, "Resonant tunneling in an AlGaAs/GaAs quantum dot as a function of magnetic field," *Phys. Rev. B* **46**, 3948 (1992).
27. H. Grabert and M. Devoret, eds., *Single Electron Tunneling*, New York: Plenum, 1991.
28. D. L. Lombardi, "Design and Self-Assembly of Conjugated Oligomers for Electronic Device Applications," Ph.D. thesis, Yale University, 1997.
29. M. Dorogi, J. Gomez, R. G. Osifchin, R. P. Andres, and R. Reifengerger, "Room temperature Coulomb blockade from a self-assembled molecular nanostructure," *Phys. Rev. B* **52**, 9071 (1995).
30. R. P. Andres, J. I. Henderson, R. G. Osifchin, *et al.*, "Coulomb staircase at room temperature in a self-assembled molecular nanostructure," *Science* **272**, 1323 (1996).
31. C. Joachim and J. K. Gimzewski, "Electronic transparency of a single C60 molecule," *Phys. Rev. Lett.* **74**, 2102 (1995).
32. C. J. Muller, J. M. van Ruitenbeek, and L. J. de Jongh, "Experimental observation of the transition from weak link to tunnel junction," *Physica C* **191**, 485 (1992).
33. C. J. Muller, J. M. Krans, T. N. Todorov, and M. A. Reed, "Quantization effects in the conductance of metallic contacts at room temperature," *Phys. Rev. B* **53**, 1022 (1996) and references therein.
34. J. M. Tour, L. Jones II, D. L. Pearson, *et al.*, "Self-assembled monolayers and multilayers of conjugated thiols, α,ω -dithiol, and thiolacetyl-containing adsorbates. Understanding attachments between potential molecular wires and gold surfaces," *J. Am. Chem. Soc.* **117**, 9529 (1995).
35. S. Datta, unpublished.
36. C. M. Fischer, M. Burghard, S. Roth, and K. von Klitzing, "Microstructured gold/Langmuir-Blodgett film/gold tunneling junctions," *Appl. Phys. Lett.* **66**, 3331 (1995).
37. H. Nejoh, "Incremental charging of a molecule at room temperature using the scanning tunneling microscope," *Nature* **353**, 640 (1991).
38. For negative bias (electrons injected from the Au bottom electrode into the conjugated molecules through the thiolates), plots of $\ln(I/V)$ vs. $1/T$ fall on

one line indicating hopping, and a similar analysis yields a hopping barrier of 0.19 eV. At present it is unclear whether the hopping is related to defects in the SAM or hopping between neighboring molecular wires.

Organic Molecular Modification of Silicon Surfaces

G. P. Lopinski, D. E. Brown, D. J. Moffatt, S. N. Patitsas,
D. D. M. Wayner, and R. A. Wolkow

*Steacie Institute for Molecular Sciences, National Research Council of Canada,
100 Sussex Drive, Ottawa, Canada.*

1. Introduction

Integration of organic molecules with existing microelectronics technology has the potential to greatly expand the capabilities of semiconductor devices. The wide range of functionality of organic molecules (molecular recognition, light emission/absorption, electron transfer, non-linear optical properties, *etc.*) is of interest for applications ranging from optoelectronic devices to biosensors. While light emitting diodes and thin film transistors based on organic molecules have existed for some time, these generally involve spin coated or evaporated films on metallic or oxide substrates. Recently there has been increasing interest in the direct modification of semiconductor surfaces via the covalent attachment of organic molecules. Novel approaches for formation of ordered molecular layers utilizing both vacuum deposition and wet chemical techniques are being developed.¹⁻¹⁰ As the chemical and physical properties of these films and nanostructures will be highly dependent on molecular structure and conformation, methods to control not only the position but also the bonding geometry of the adsorbed molecule will be required. Progress in this area requires a detailed understanding of how organic molecules interact with semiconductor surfaces including knowledge of bonding configurations as well as the dynamical processes (*i.e.* reactions, diffusion) involved in adsorption.

In this article we review some recent results from our scanning tunneling microscopy (STM) investigations of organic molecule adsorption on silicon surfaces. The examples presented here serve to summarize the current level of understanding of organic molecule adsorption on semiconductor surfaces and illustrate current capabilities for determining as well as controlling the position and bonding configuration of adsorbed molecules. Possible obstacles to progress in this area such as limited surface diffusion (hindering self-assembly, formation of ordered layers) and electron-induced Si-C bond breaking (a possible degradation mechanism in molecular devices) are also identified.

2. Imaging bonding geometries

The scanning tunneling microscope is a uniquely powerful tool with the ability to image and manipulate individual atoms and molecules. However, while STM

images can show approximately where molecular adsorption has occurred, it is in general difficult to extract details of the adsorbate bonding geometry. Interpretation of STM images is complicated as they involve a convolution of topography and electronic state information. Furthermore, covalent bonding with the dangling bonds of the semiconductor surface will drastically alter the geometry and electronic states of the adsorbed molecule. As a result, some form of theoretical modeling is necessary to understand the observed STM features. We have used a silicon cluster to model the surface, obtaining optimized bonding geometries and adsorption energies with a variety of quantum chemistry computational methods including the semi-empirical AM1 method,¹¹ as well as Hartree-Fock (HF) and density functional (B3LYP) methods.¹² This approach has previously been employed by Raghavachari *et al.* to successfully model the interaction of hydrogen and water with silicon surfaces.^{13,14} Charge density iso-surfaces, constructed from sums of all the energetically accessible molecular orbitals are also obtained. Within the Tersoff-Hamman approximation,¹⁵ these will simulate constant current STM images. Comparison of these synthesized images with those observed experimentally allows specific adsorbate bonding configurations to be identified.

The Si(100) surface consists of rows of silicon dimers, with each dimer having two dangling bonds. A clean surface can be prepared by heating to 1250 °C under ultrahigh vacuum conditions ($<10^{-10}$ torr) to desorb the native oxide layer. While saturated hydrocarbons do not react with this surface, unsaturated molecules such as alkenes and alkynes have been shown to react readily with the available dangling bonds.^{16,17} Figure 1 shows an STM image of trans-2-butene (C_4H_8) adsorbed onto Si(100) by controllably leaking the gaseous molecules into the vacuum chamber. The molecules are imaged as paired protrusions, centered along a dimer row. Comparison with the calculated bonding geometry (HF/3-21G*) shows that these protrusions arise from the two methyl groups on either end of the molecule. In this configuration, the two central carbons, sp^2 before adsorption, become rehybridized to sp^3 by forming covalent bonds with the dangling bonds of a silicon dimer. Similar bonding geometries are observed for other simple alkenes such as ethylene and propylene. The simulated STM image is seen to show good agreement with the experiment. The methyl groups were not expected to give rise to maxima in the images due to the large energy gap associated with saturated hydrocarbons. However, the calculations show that covalent bonding to the silicon modifies the orbitals associated with the molecule, inducing a small electronic density of states on the methyl groups for energies in the vicinity of the conduction band edge (with a bias voltage of +2 V, states within 1.5 eV of the band edge are accessible for tunneling). As the tunneling current is an exponential function of distance even a small density of states that is physically closer to the tip will contribute significantly to the image. We note that the measured "height" of the methyl groups relative to the clean Si dimers is only 0.6 Å at +2 V, much less than the actual distance above the surface of 2.5 Å.

The ability to resolve the position of individual methyl groups allows the geometric configuration (cis or trans) of the adsorbed alkenes to be determined.¹⁰ For cis-2-butene, in which the methyl groups are on the same side of the molecule, the paired protrusions lie perpendicular to the dimer row direction, not at 30° as

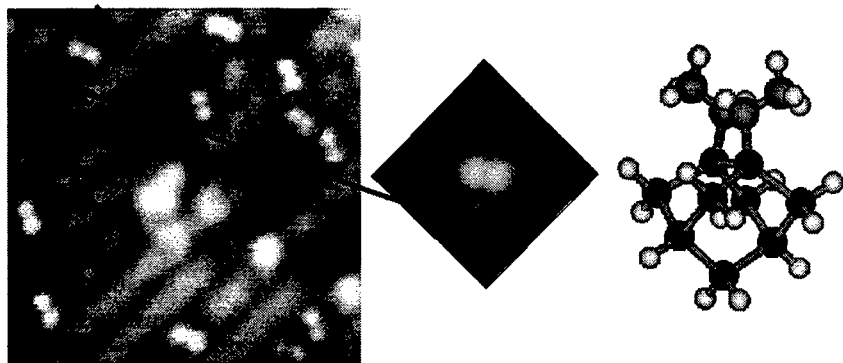


Figure 1. Unoccupied state STM image of trans-2-butene/Si(100) ($75 \times 75 \text{ \AA}$, $V_s = +2 \text{ V}$, $I = 40 \text{ pA}$), together with a simulated image for a five Si dimer cluster and optimized bonding geometry for a single dimer cluster. The circle denotes an adsorbed cis-2-butene molecule. The darkest and brightest features correspond to defects.

observed for the trans isomer. As a result the STM can be used to probe the degree to which adsorption of a given molecule is stereoselective (i.e. retains its geometric configuration). In Fig. 1 a single cis impurity is observed even though the trans isomer was introduced into the chamber (the level of cis impurity in the gas is $\leq 0.3\%$). Analysis of many adsorbed molecules indicates the concentration of the adsorbed cis configuration is $2.1 \pm 0.7\%$. This result indicates that while the reaction of 2-butenes with the Si(100) surface is stereoselective, in agreement with previous desorption studies,¹⁸ there is a small probability of isomerization upon adsorption.

The adsorption of alkenes is a rather simple case as the molecules are only observed to react in a single manner. Is the combination of the STM imaging together with cluster calculations useful for determining the bonding geometry of more complex molecules? To answer this question we have studied the adsorption of benzene, the prototypical aromatic molecule. As such it serves as a useful starting point for understanding the adsorption of larger conjugated molecules such as pentacene and anthracene which form organic semiconductors with high field-effect mobilities. Furthermore, the size of benzene is similar to the distance between Si dimers along a row, suggesting that bridging between these dimers may be possible. Previous experimental work has shown that benzene adsorbs molecularly, with significant rehybridization of some of the carbon atoms.¹⁹ Calculations using semi-empirical methods differ on whether bonding to a single dimer²⁰ or a bridging geometry²¹ yields the most stable adsorption site.

STM images of benzene/Si(100) indicate the presence of multiple bonding configurations as three different types of features due to benzene molecules are observed. With the aid of the cluster calculations the geometry giving rise to each of these features can be identified as illustrated in Fig. 2. For all the calculated structures the molecule loses its aromatic character, becoming significantly

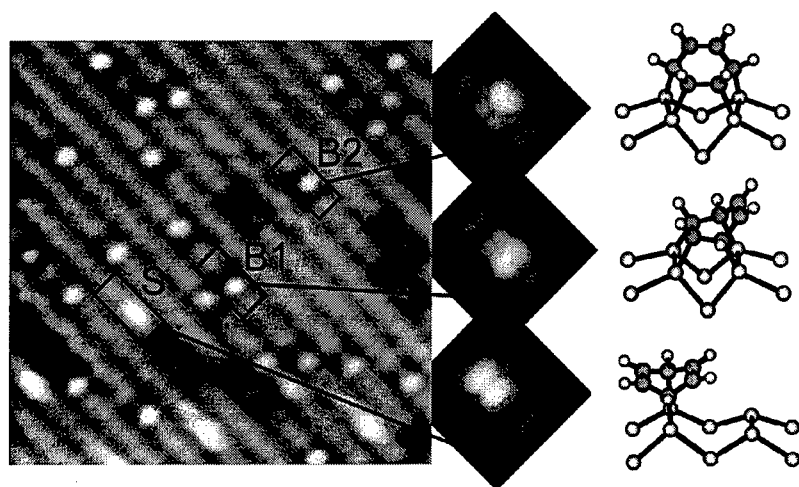


Figure 2. Occupied state image of benzene/Si(100) ($100 \times 100 \text{ \AA}$, $V_s = -1.5 \text{ V}$, $I = 40 \text{ pA}$) together with synthesized images (4 dimer cluster) and calculated structures (2 dimer cluster), from top to bottom; twisted bridge, tight bridge and single dimer.

distorted due to Si-C bond formation and rehybridization. The brightest protrusions (labeled S), centered over a single silicon dimer, are seen to arise from molecules bound to a single dimer in a butterfly configuration. The protrusion is due to the two double bonds that remain on the adsorbed molecule. The features labeled B1 and B2 both occupy two Si dimers and are assigned to two different bridging configurations. These geometries leave one double bonded unit remaining in the molecule, oriented parallel (B1) or perpendicular (B2) to the dimers, giving rise to the bright feature within the darkened area. Both of these bridging geometries require significant distortion of the Si lattice. The B1 feature requires a considerable shortening of the distance between dimers ("tight" bridge) while a twist of the dimers in the plane of the surface is required to accommodate the B2 configuration ("twisted" bridge). Adsorption energies, calculated at the B3LYP/6-31G* level, indicate the tight bridge to be considerably more stable (1.49 eV) than the twisted bridge (0.91 eV) or the single dimer geometry (0.88 eV). Infrared spectroscopy of the C-H stretch modes using the multiple internal reflection geometry indicates rehybridization of several of the carbons to sp^3 with some remaining sp^2 . Comparison of the measured spectrum with calculations yields evidence for all three of the configurations shown above.⁹

Consideration of the calculated adsorption energies suggests that the single dimer state should relax to the more stable tight bridge state. Furthermore, from the geometries depicted in Fig. 2 it is evident that such a relaxation would involve simply leaning over to one side and forming two additional Si-C bonds. In fact, while images taken shortly after dosing show most of the molecules in the single dimer state, time sequences of images over two hours reveal a slow conversion to

the tight bridge. The activation barrier for this conversion is determined to be 0.94 eV. As for the twisted bridge (B2), adsorption in this configuration is found to be a minority species occurring exclusively at a particular defect on the Si(100) surface known as type C. At these type C defects, thought to involve adjacent dimers buckled in the same direction as a result of a subsurface substitutional impurity, the dimers may be predistorted to accommodate the twisted bridge geometry. We note that, unlike the alkenes, which become completely saturated upon adsorption, all the observed benzene configurations have at least one double bond remaining on the molecule. This double bond presents an opportunity to carry out further chemistry such as a Diels-Alder addition, using the covalently attached benzene as a foothold on the surface.

In summary, these two examples indicate that the combination of STM and theoretical modeling using a cluster/quantum chemistry approach is quite effective in determining bonding geometries of organic molecules on silicon surfaces. This conclusion remains true even for more complex systems such as benzene/Si(100), for which adsorption leads to significant distortions of both the molecule and the substrate as well as the occurrence of multiple bonding geometries.

3. Tip-induced Si-C bond breaking

Fabrication of devices based on organic molecules will require patterning of these layers. The STM tip can be used for atomic resolution lithography as demonstrated by desorption of hydrogen atoms from the H-terminated Si(100) surface.²² Two distinct mechanisms of tip induced hydrogen desorption have been identified:²²⁻²⁴ electron-induced electronic excitation and vibrational heating. As the H-terminated surface is not reactive towards organic molecules under vacuum conditions, adsorption will only occur where hydrogen has been removed to create Si dangling bonds. Spatially selective adsorption of organic molecules using this approach has been used to create 250 X 150 Å regions of norbornadiene/Si(100).²¹

We have demonstrated that it is also possible to break Si-C bonds with a tip induced process, offering a strategy for the direct patterning of organic films. In particular, by imaging benzene/Si(100) at slightly elevated tip voltages, electrons from the STM can locally induce desorption, diffusion, and conversion between two different chemisorbed states. Figure 3 shows two images of the same area of a Si(100) surface with a low coverage of adsorbed benzene before and after scanning at a sample bias of -3 V. Immediately apparent is the larger fraction of bright protrusions, characteristic of the single dimer state, in the image on the right. Further comparison of the images reveals that all the molecules have moved except for the circled ones. At these locations, tight bridge features have converted directly to single dimer protrusions at the same position. This conversion is the reverse of the thermally induced binding state conversion discussed in the previous section. Inspection of the geometries accompanying Fig. 3 indicates that conversion from the tight bridge to single dimer state involves breaking two Si-C bonds. The observation of considerable diffusion is also surprising as imaging at lower bias voltages has shown that there is no thermally induced diffusion at room

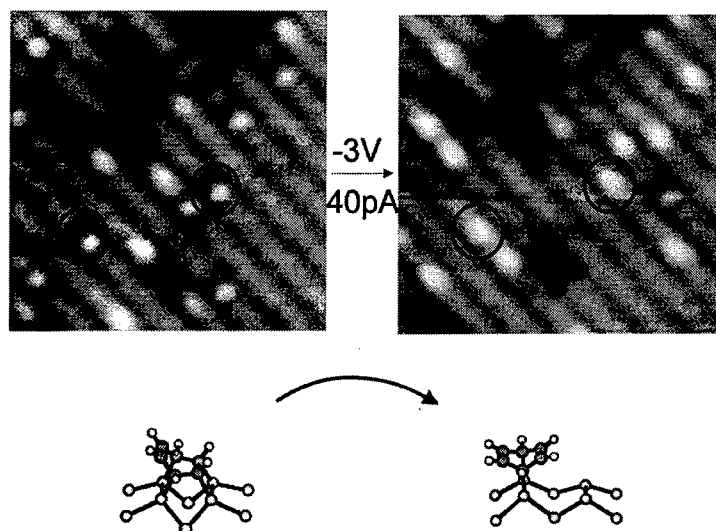


Figure 3. STM images of benzene/Si(100) ($75 \times 75 \text{ \AA}$, $V_s = -1.5 \text{ V}$, $I = 40 \text{ pA}$) showing the effects of scanning at elevated bias voltage (-3 V). The circled regions indicate two induced sites where the tip has induced conversion from the tight bridge to single dimer configurations as shown schematically below the image.

temperature. Analysis of the tip-induced diffusion events indicates that the average jump length is $10 \pm 5 \text{ \AA}$, approximately 2.5 dimer units, with no apparent preference for diffusion along or perpendicular to a dimer row. Comparison of the images also indicates there are two fewer molecules in the right hand image. While these could have simply diffused out of the image, larger area scans confirm that the tip is inducing desorption. Both desorption and diffusion require breaking of all four Si-C bonds. Diffusion, however, requires that the molecule remains bound to the surface, suggesting the existence of a mobile, metastable state. A molecule excited into this weakly bound state by the tip can diffuse along the surface and either desorb or return to the more stable covalently bound state. In the present case benzene molecules re-bonding with the surface will have to pass through the single dimer state before being able to bridge, accounting for the increased number of single dimer molecules observed after tip-induced diffusion. As discussed in the next section, a weakly bound metastable state has in fact been observed directly for benzene/Si(111) using low temperature STM.

Regarding the mechanism for the observed Si-C bond breaking, the threshold for this process is around 2.3 eV and the probability is roughly linear in the applied current. No tip-induced desorption or diffusion has been observed for the alkenes, which become totally saturated upon chemisorption, suggesting that the π electrons remaining on the rehybridized benzene play a crucial role. The tip breaking occurs during tunneling of occupied states (i.e. electrons moving from sample to tip). While assignment of a mechanism is somewhat still speculative,

the facts suggest desorption via a positive ion resonance created by removal of an electron from a π state of the molecule. This case differs from the one of tip-induced hydrogen desorption, which is typically observed at positive sample bias (electron flow to the sample). However, H-desorption at negative bias via a hole resonance was reported recently.²⁶ Besides being of interest for patterning, these tip-induced reactions can model electron stimulated effects in devices. The low voltage, low current electron induced Si-C bond breaking observed here suggests a possible degradation problem for devices based on unsaturated molecules.

4. Precursor-mediated diffusion

Schemes for making molecular devices based on the concept of self-assembly require substantial diffusion of adsorbed molecules. However, for organic molecules covalently bound to semiconductor surfaces, diffusion barriers are found to be rather large. For several different alkenes and benzene on Si(100) we found no evidence of diffusion at room temperature. This finding is in agreement with previous studies of benzene/Si(111) in which diffusion was observed only for temperatures close to the desorption temperature.²⁷ When diffusion did occur, the average jump length corresponded to several lattice sites, similar to the tip-induced diffusion for benzene/Si(100), which also involved multiple site hops. These observations are in contrast to the usual situation for adsorbates on metal surfaces where the diffusion barrier is generally much smaller ($\sim 10\%$ of the desorption energy) and usually takes place via single site hopping. The present observations can be explained if diffusion occurs via a weakly bound, mobile "precursor" state, as suggested in the previous section.

Consider the potential energy diagram in Fig. 4(a) for an adsorbate approaching a surface. The system passes through a weakly bound (physisorbed) state on its way to the covalently bound (chemisorbed) state. While in the chemisorbed state there is a high barrier to lateral motion as this involves breaking of covalent bonds, in the physisorbed state the molecule is expected to be highly mobile. Thus diffusion can occur by first surmounting the barrier from chemisorbed to physisorbed states. As can be seen from Fig. 4, this model implies the barrier to diffusion is similar to that for desorption. In this model the diffusion length will be determined by the diffusion barrier in the physisorbed state as well as the time-scale for rechemisorption. If the diffusion barrier is significantly smaller than the barrier for rechemisorption, multiple site hopping will result.

Using a variable temperature STM has made it possible to trap molecules in this precursor state, providing the first direct observation of this evasive yet probably common entity.²⁸ Figure 4 shows benzene adsorbed on the Si(111) surface at two different temperatures. For adsorption in the 100–300 K range, benzene appears as a darkening of the Si adatoms of the (7 \times 7) reconstruction, as in Fig. 4(b). The appearance of benzene as "missing" adatoms is understood in terms of covalent bonding quenching the adatom dangling bonds, which dominate the clean surface image. For adsorption at 78 K, the image is dramatically different, the honeycomb pattern. These protrusions can be associated with

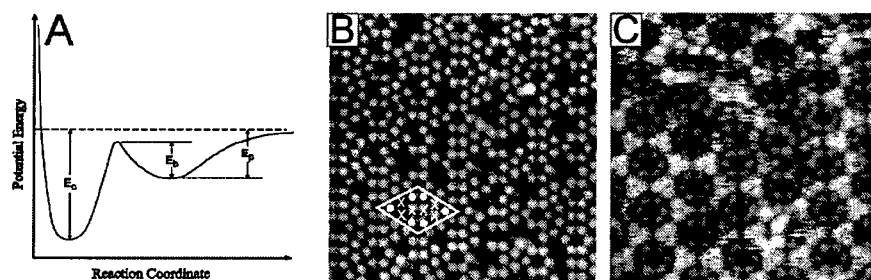


Figure 4. A) Schematic potential energy diagram describing precursor mediated adsorption. B) STM image of benzene/Si(111) at 120 K. The white diamond marks the unit cell of the (7X7) reconstruction with the six corner and six center adatoms denoted by white circles and X's, respectively. C) STM image of benzene/Si(111) at 78 K. Both images are 140 X 140 Å areas ($V_s = +1.5$ V, $I = 100$ pA).

physisorbed benzene; the "fuzziness" of the image is associated with motion of the molecule on the time scale of scanning. The honeycomb pattern indicates that the molecules are not completely delocalized across the surface but prefer to spend time over certain sites, in particular the center adatoms. In images at 95 K, this pattern becomes more noisy and less distinct. Above 110 K, the precursor is too mobile to be observed. From these observations the diffusion barrier in the precursor state can be estimated to be ~ 0.2 eV. While observation of physisorbed molecules is not in itself remarkable, the precursor state is unique in that it occurs over a site where chemisorption can and will occur. In fact, sequences of images reveal decay of some of the molecules from the precursor to chemisorbed states. By monitoring the rate of this process the barrier to chemisorption, E_b , can be estimated to be 0.3 eV. Using the measured barriers for diffusion and re-chemisorption the expected diffusion length at room temperature can be calculated to be 27 Å, close to the experimental value of 23 ± 1 Å.

Although the precursor state observed in Fig. 4(c) was observed directly upon dosing at 78 K, the STM tip can also be used to prepare this state at will, starting from the chemisorbed benzene. Conditions for dislodging molecules from the chemisorbed into the physisorbed state occur under similar conditions (-3 V, 100 pA) as required for the Si-C bond breaking observed for benzene/Si(100), suggesting a similar mechanism may be operative. Both these observations indicate that tip-induced population of the physisorbed precursor state offers a possible route for facilitating surface diffusion.

5. Wet chemical modification

While UHV offers a well-controlled environment for studying organic adsorption and making ordered films, it is ultimately desirable to do modifications via wet-

chemical methods in order to simplify processing and lower cost. In order to form well-ordered, covalently bound layers of organic molecules it is necessary to start from an atomically flat, oxide free surface. Chabal and Higashi showed that it is possible to form monohydride terminated Si(111) surfaces by etching with ammonium fluoride.²⁹ Chidsey has used this surface as a starting point for further chemistry, producing densely packed alkyl monolayers from alkenes using a radical initiated reaction.^{1,2} Lewis has prepared alkyl terminated Si surfaces via a two step process involving replacement of the hydrogen with chlorine followed by reaction with alkyl-Li or alkyl-Grignard.³ Electrochemical characterization indicates that these alkyl terminated surfaces have a rather low density of surface recombination sites (comparable to that of the H-terminated surface), corresponding to one electrically active defect every 10^5 atoms.³ While these insulating alkyl chains are of interest as passivating layers, molecular devices will also involve conducting molecules. Oligothiophenes are conjugated molecules that are among the organic semiconductors of choice for thin film transistors and may be interesting candidates for building efficient organic-based lasers.³⁰ Covalent attachment of thiophene and terthiophene molecules on Si(111) has recently been demonstrated by bromination followed by reaction with thienyllithium.⁸ We are currently working towards STM imaging and characterizing the electronic properties of these modified surfaces.

6. Concluding remarks

Covalent bonding of organic molecules to semiconductor surfaces has the potential to lead to novel hybrid devices with expanded functionality. The understanding of organic chemistry on silicon surfaces, necessary for further progress in this area, is still being developed. We have shown that the combination of STM and quantum chemistry calculations, using a cluster model for the surface, is a useful approach for determining the bonding geometry of adsorbed organic molecules. The interaction of small molecules such as alkenes and benzene with the Si(100) surface is now fairly well understood, and can be used as the starting point for attachment of more complex molecules with interesting molecular recognition or optoelectronic properties. Methods for control of the adsorbate bonding geometry via tip-induced modifications are also being explored. Tip-induced processes could prove useful for inducing diffusion of the adsorbed molecules. The developing wet chemical approaches for organic modification appear promising. Results obtained thus far indicate that these methods have tremendous potential for producing well-ordered layers with good electrical characteristics via relatively simple processing.

References

1. M. R. Lindford and C. E. D. Chidsey, "Alkyl monolayers covalently bound to silicon surfaces," *J. Am. Chem. Soc.* **115**, 12631 (1993).

2. M. R. Lindford, P. Fenter, P. M. Eisenberger, and C. E. D. Chidsey, "Alkyl monolayers on silicon prepared from 1-alkenes and hydrogen-terminated silicon," *J. Am. Chem. Soc.* **117**, 3145 (1995).
3. A. Banshal, X. Li, I. Lauermann, *et al.*, "Alkylation of Si surfaces using a two-step halogenation/Grignard route," *J. Am. Chem. Soc.* **118**, 7225 (1996).
4. R. J. Hamers, J. S. Hovis, S. Lee, H. Liu, and J. Shan, "Formation of ordered, anisotropic organic monolayers on the Si(001) surface," *J. Phys. Chem.* **101**, 1489 (1997).
5. J. S. Hovis and R. J. Hamers, "Structure and bonding of ordered organic monolayers of 1,5 cyclooctadiene on the silicon(001) surface," *J. Phys. Chem.* **101**, 9581 (1997).
6. R. Konecny and D. J. Doren, "Theoretical prediction of a facile Diels-Alder reaction on the Si(100)-2X1 Surface," *J. Am. Chem. Soc.* **119**, 11098 (1997).
7. A. V. Teplyakov, M. J. Kong, and S. F. Bent, "Vibrational spectroscopic studies of Diels-Alder reactions with the Si(100)-2X1 surface as a dienophile," *J. Am. Chem. Soc.* **119**, 11100 (1997).
8. J. He, S. N. Patitsas, K. F. Preston, R. A. Wolkow, and D. D. M. Wayner, "Covalent bonding of thiophenes to Si(111) by a halogenation/thienylation route," *Chem. Phys. Lett.* **286**, 508 (1998).
9. G. P. Lopinski, T. M. Fortier, D. J. Moffatt, and R. A. Wolkow, "Multiple bonding geometries and binding state conversion of benzene/Si(100)," *J. Vac. Sci. Technol. A* **16**, 1037 (1998).
10. G. P. Lopinski, D. J. Moffatt, D. D. M. Wayner, and R. A. Wolkow, "Determination of the absolute chirality of individual adsorbed molecules using the scanning tunneling microscope," *Nature* **392**, 909 (1998).
11. M. J. S. Dewar and W. Thiel, "Ground states of molecules. The MNDO method. Approximations and parameters," *J. Am. Chem. Soc.* **99**, 4499 (1977);
M. J. S. Dewar, E. G. Zoebisch, and E. F. Healy, "AM1: A new general purpose quantum mechanical molecular model," *J. Am. Chem. Soc.* **107**, 3902 (1985).
12. M. J. Frisch *et al.*, *Gaussian 94* (Revision D.1), Gaussian Inc., Pittsburgh, PA (1995).
13. Y. J. Chabal and K. Raghavachari, "Surface infrared study of Si(100)-(2X1)H," *Phys. Rev. Lett.* **53**, 282 (1984).
14. M. K. Weldon, B. B. Stefanov, K. Raghavachari, and Y. J. Chabal, "Initial H₂O-induced oxidation of Si(100)-(2X1)," *Phys. Rev. Lett.* **79**, 2851 (1997).
15. J. Tersoff and D. R. Hamman, "Theory and application for the scanning tunneling microscope," *Phys. Rev. Lett.* **50**, 1998 (1983).
16. M. J. Bozack, P. A. Taylor, W. J. Choyke and J. T. Yates, Jr., "Chemical activity of the C=C double bond on silicon surfaces," *Surf. Sci.* **177**, L933 (1986).
17. C. C. Cheng, R. M. Wallace, P. A. Taylor, W. J. Choyke, and J. T. Yates Jr., "Direct determination of absolute monolayer coverages of chemisorbed C₂H₂ and C₂H₄ on Si(100)," *J. Appl. Phys.* **67**, 3693 (1990).

18. M. Kiskinova and J. T. Yates, Jr., "Observation of steric conformational effects in hydrocarbon adsorption and decomposition: cis- and trans-butene-2 on Si(100)-(2X1)," *Surf. Sci.* **325**, 1 (1995).
19. Y. Taguchi, M. Fujisawa, T. Takaoka, T. Okada, and M. Nishijima, "Adsorbed state of benzene on the Si(100) surface: thermal desorption and electron energy loss spectroscopy studies," *J. Chem Phys.* **95**, 6870 (1991).
20. B. I. Craig, "A theoretical examination of the chemisorption of benzene on Si(100)-(2X1)," *Surf. Sci.* **280**, L279 (1993).
21. H. D. Joeng, S. Ryu, Y. S. Lee, and S. Kim, "A semi-empirical study of the chemisorbed state of benzene on Si(100)-(2X1)," *Surf. Sci.* **344**, L1226 (1995).
22. T. C. Shen, C. Wang, G. C. Abeln, *et al.*, "Atomic-scale desorption through electronic and vibrational excitation mechanisms," *Science* **268**, 1590 (1995).
23. T. C. Shen and Ph. Avouris, "Electron stimulated desorption induced by the scanning tunneling microscope," *Surf. Sci.* **390**, 35 (1997).
24. E. T. Foley, A. F. Kam, J. W. Lyding, and Ph. Avouris, "Cryogenic UHV-STM study of hydrogen and deuterium desorption from Si(100)," *Phys. Rev. Lett.* **80**, 1336 (1998).
25. G. C. Abeln, S. Y. Lee, J. W. Lyding, D. S. Thompson, and J. S. Moore, "Nanopatterning organic monolayers on Si(100) by selective chemisorption of norbornadiene," *Appl. Phys. Lett.* **70**, 2747 (1997).
26. K. Stokbro, C. Thirstrup, M. Sakurai, *et al.*, "STM-induced hydrogen desorption via a hole resonance," *Phys. Rev. Lett.* **80**, 2618 (1998).
27. R. A. Wolkow and D. J. Moffatt, "The frustrated motion of benzene on the surface of Si(111)," *J. Chem Phys.* **103**, 10696 (1995).
28. D. E. Brown, D. J. Moffatt, and R. A. Wolkow, "Isolation of an intrinsic precursor to molecular chemisorption," *Science* **279**, 542 (1998).
29. G. S. Higashi, Y. J. Chabal, G. W. Trucks, and K. Raghavachari, "Ideal hydrogen termination of the Si(111) surface," *Appl. Phys. Lett.* **56**, 656 (1990).
30. F. Garnier, G. Horowitz, P. Valat, F. Kouki, and V. Wintgens, "The four-level stimulated emission in sexithiophene single crystals," *Appl. Phys. Lett.* **72**, 2087 (1998).

4 The Message is the Media: Storage Materials and Technologies

This chapter addresses semiconductor memories and magnetic storage media. According to some projections, in the near future "sweet" memories are poised to capture over 50% of the total IC market. At the same time, the magnetic storage market is huge and growing. It looks like the amount of money we shall be spending on remembering things will far outstrip all our outlays on producing things to remember. Moreover, dynamic memories that require constant refreshing, such as the DRAM that has served as the major technology driver in the semiconductor industry, were seen by the great majority of Embiez attendees to be yielding to more permanent "nonvolatile" memories.

The chapter opens with a retrospective by Simon Sze, one of the co-inventors of the original nonvolatile "floating-gate" memory concept. Professor Sze is known universally both for his pioneering contributions and for his masterful books that have been for years the primary source of knowledge for numerous research workers. When Sze predicts, the world listens, and here he predicts that nonvolatile memories will be integrated in the widest variety of applications, including mass storage and embedded systems, and that, moreover, they will become one of the most significant technology drivers, culminating in the development of single-electron nonvolatile memory cells. Interestingly, similar predictions were heard from the single-electron quarters! Konstantin Likharev, one of the pioneers of the single-electron device concept, described the ultra-dense nonvolatile memory chip as the most significant, *if not the only* practical application of single electronics.

Magnetic storage is another versatile technology that is expected to survive and flourish in the nanoelectronic age. Prophecies of hybrid systems, so common in this book, extend to the marriage of magnetic storage with random access memories, advocated by Arto Nurmikko and Herbert Goronkin. "When in doubt, think hybrid", they admonish us from their joint academic/industrial perspective.

Volatile memory adherents were a minority at Embiez. Fortunately, science is not decided by a majority vote. When told of the new book, *100 Scientists against Relativity*, Einstein famously remarked that "if I were wrong, one would be enough". An imaginative vertically stacked SRAM structure, discussed by Marco Mastrapasqua and co-authors in this chapter, may well become a prototypical way for 3D circuit integration.

Evolution of Nonvolatile Semiconductor Memory: From Floating-Gate Concept to Single-Electron Memory Cell

S. M. Sze

*National Chiao Tung University and National Nano Device Laboratories, Hsinchu,
Taiwan, R.O.C.*

1. Introduction

Semiconductor memories constitute about 30% of the world integrated circuit (IC) market.¹ In ten years, they will capture the largest market share with over 50% of the total IC sales. Therefore, semiconductor memories will stop being looked at as an auxiliary to the microprocessors (or other logic circuits) and microprocessors will start being looked at as aids to the optimized and predominant memories on IC chips.

Semiconductor memories can be broadly classified into two groups: volatile and nonvolatile.² Volatile memories such as DRAM (dynamic random access memory) and SRAM (static random access memory) lose the stored information once the power supply is switched off. Nonvolatile memories, on the other hand, can retain the stored information.

The nonvolatile semiconductor memory (NVSM) group includes the following members:³ (1) ROM (read-only memory) with data permanently written during manufacturing; (2) PROM (programmable ROM) with data that can be written only once using a fuse or antifuse process; (3) floating-gate memory; (4) silicon-nitride memory; (5) FRAM (ferroelectric RAM); and (6) MRAM (magnetoresistive RAM).

In order to compare the characteristics and performance of these memories, we first define twelve attributes of an ideal memory: (1) nonvolatility with long retention (i.e., the ability to retain data without power over a long period of time, typically over 10 years); (2) high density (i.e., small cell area); (3) low power consumption; (4) in-system rewritability; (5) bit alterability; (6) fast read/write; (7) high endurance (i.e., the ability to maintain stored information after repeated erase/program/read cycling); (8) low cost; (9) single power supply; (10) high scalability (i.e., cell size can be reduced with the minimum feature length); (11) ruggedness; and (12) high integrability with silicon IC technology.

A comparison³ of various memories based on the aforementioned attributes is shown in Table 1. The ROM and PROM are not included, since they can not be re-programmed. Neither is silicon-nitride memory because of its low endurance and low retention. The floating-gate memory can be subdivided into three categories: flash memory (see Sec. 4 below), EEPROM (electrically erasable programmable

Device Attribute	DRAM	SRAM	Flash	EEPROM	UV- EPROM	FRAM	Hard Disk	Floppy Disk
1. Nonvolatility	-	-	✓	✓	✓	✓	✓	✓
2. High density	✓	-	✓	-	✓	✓	✓	-
3. Low power	-	-	✓	✓	✓	✓	-	-
4. In-system rewritability	✓	✓	✓	✓	-	✓	✓	✓
5. Bit alterability	✓	✓	-	✓	-	✓	✓	✓
6. Fast read/write	✓	✓	✓	✓	✓	✓	-	-
7. High endurance	✓	✓	✓	✓	-	✓	✓	✓
8. Low cost	✓	-	✓	-	✓	-	✓	✓
9. Single-power supply	✓	✓	✓	✓	-	✓	✓	✓
10. Highly-Scalable	-	-	✓	-	✓	✓	-	-
11. Ruggedness	✓	✓	✓	✓	✓	✓	-	-
12. Highly integrable	✓	✓	✓	✓	✓	-	-	-
Total Desirable Attributes	9	7	11	9	8	10	7	6

Table 1. Comparison of Memory Attributes.³

ROM), and UV-EPROM (ultraviolet erasable programmable ROM).

As can be seen from Table 1, no memory has all the attributes of an ideal memory. For example, DRAM suffers from relatively high power consumption and a non-scalable storage capacitor in addition to its volatility. SRAM suffers from low density and volatility. FRAM and MRAM are still in their initial development stage and the main concerns are their process uniformity and compatibility with silicon IC processing. We have also included the hard and floppy disks, which suffer from high power consumption and low ruggedness. The memory that has the most desirable attributes is the flash memory in the floating-gate memory family. In the subsequent discussion, we will concentrate on the dominant NVSM — the floating-gate memory.

The historical developments of the floating gate memory will be presented in Sections 2—5; their major applications, especially for portable electronic systems, will be discussed in Sections 6—8; followed by a brief conclusion.

2. History of the floating-gate concept

The NVSM was first proposed⁴ by Kahng and Sze in 1967. The proposal introduced the floating gate concept for charge storage and nonlinear transport processes for programming and erasing. From a historical perspective, the 1967 proposal recognized for the first time the possibility of rewritable NVSM devices.

Figure 1 shows the cross-sectional view of the first NVSM with a floating gate M(1). The basic operations of programming (e.g., Fowler-Nordheim tunneling), storage, and erase (e.g., reverse F-N tunneling) are shown in Fig. 2. If

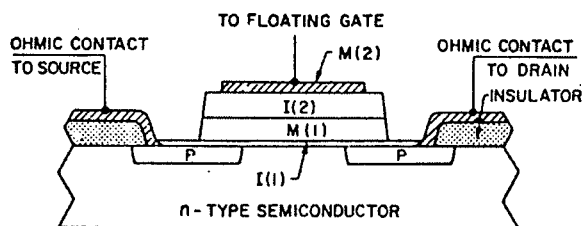


Figure 1. Cross-sectional view of the first proposed nonvolatile semiconductor memory with a floating gate⁴ in 1967.

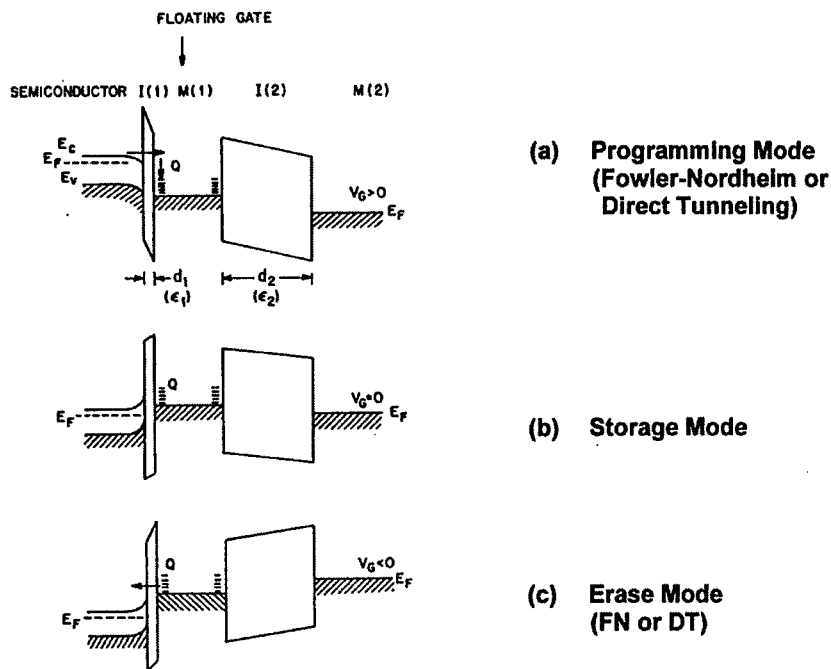


Figure 2. Programming, storage, and erase modes for the floating-gate memory.⁴

the insulators I(1) and I(2) are sufficiently thick, the charge in the floating gate can be stored for a long time.

An experimental device was also made using SiO_2 as I(1), Zr as M(1), and ZrO_2 as I(2). When a voltage pulse of 50 V with a pulse duration of 0.5 μs was applied to the control gate, M(2), about 10^{12} electrons/ cm^2 were transported to and stored in the floating gate. The stored charge caused a large threshold voltage shift, and the device was "on" with a channel current of 0.25 mA. When a large negative pulse was applied to the control gate, the stored charge was depleted and the device

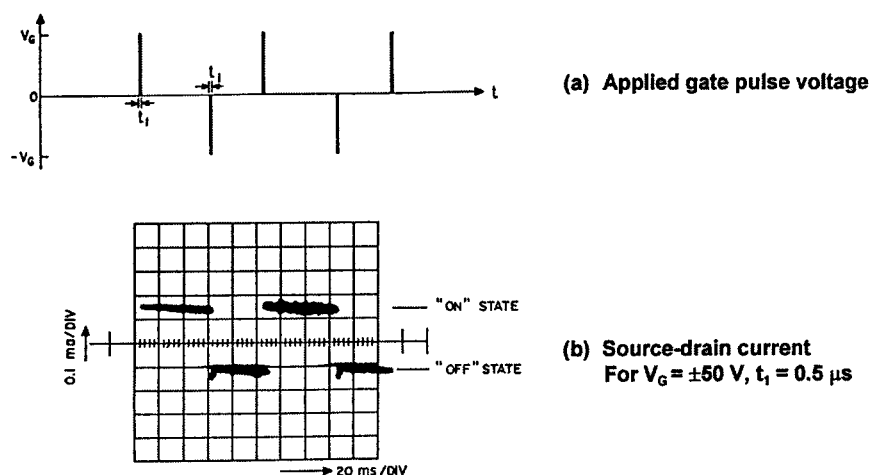


Figure 3. First demonstration of an electrically erasable programmable read-only memory (EEPROM).⁴

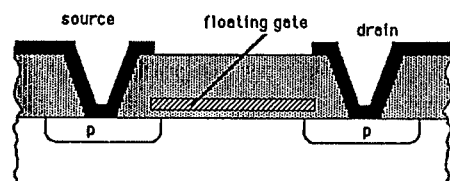
turned "off" as shown in Fig. 3. The slope of the "on" state indicates that this NVSM can store information (i.e., charge) for a few hundred milliseconds. This result can be considered as the first demonstration of EEPROM operation.

3. Early device structures

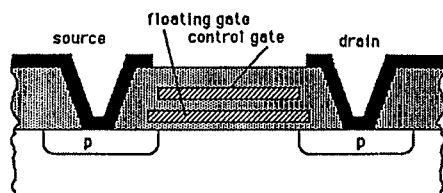
In 1971, Frohman-Bentchkowsky developed FAMOS⁵ (floating-gate avalanche-injection MOS) shown in Fig. 4 (top). This structure is an EPROM and has a floating gate but no control gate. The stored charge in the floating gate came from injection of hot electrons generated at the avalanche region near the drain. Since FAMOS has no control gate, the stored charge can not be erased electrically. One has to use ultraviolet light to do the erasing.

In 1976, Iizuka and his coworkers studied the SAMOS⁶ (stacked-gate avalanche-injection MOS) shown in Fig. 4 (bottom). This structure is an EEPROM, and it looks essentially the same as that in Fig. 1. However, the injection mechanism is due to avalanche instead of Fowler-Nordheim tunneling. Because of the relatively thick oxide between the floating gate and the channel, a substantial improvement of the retention time was obtained. However, for a typical EEPROM operation, we need two devices per cell, i.e., an EEPROM and a selection MOSFET. Therefore, the cell size is relatively large.

In 1984, Masuoka and his coworkers developed the flash memory.⁷ In its erase operation, a whole block is erased, thus the name "flash". The top and cross-sectional view of a flash memory are shown in Fig. 5. Since there is only one device per cell, flash memory has the advantages of higher density, lower cost, and higher scalability compared to the EEPROM.



In 1971
FAMOS (Floating Gate
Avalanche Injection MOS)



In 1976
SAMOS (Stacked Gate
Avalanche Injection MOS)

Figure 4. Two early floating-gate memories.^{5,6}

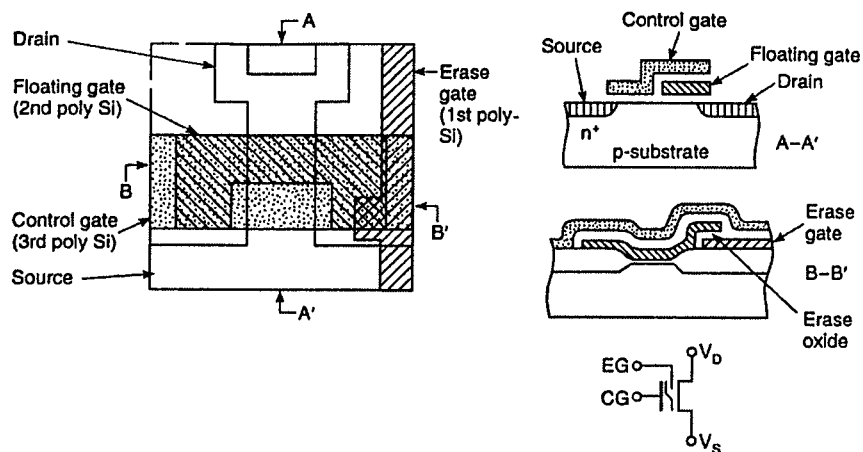


Figure 5. Flash memory⁷ proposed in 1984. In the erase operation, a whole block is erased, thus the name "flash".

All early NVSMs were not very reliable. They all suffered from low endurance and poor retention. This is expected, since the injection (or reverse injection to remove the stored charge) mechanisms of the avalanche process and Fowler-Nordheim tunneling generally cause reliability problems in conventional MOSFETs. For example, hot-electron trapping or time-dependent dielectric breakdown (TDDB) in the oxide may induce leakage currents, which in turn, may result in poor endurance and low retention.

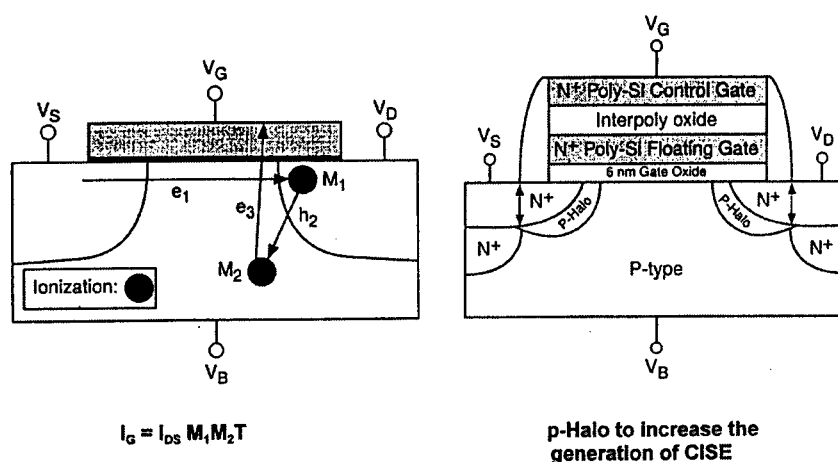


Figure 6. Schematic diagram of the channel-induced substrate electron injection process.⁸ Gate current I_G is determined by the multiplication M_1 , creating holes that ionize in the substrate with multiplication M_2 , and then tunnel into the gate with probability T .

To improve the reliability, the programming voltage for Fowler-Nordheim tunneling can be reduced by using a thinner oxide between the floating gate and the channel. The programming voltage for hot-electron injection can also be reduced by using the channel-induced substrate electron injection (CISEI)⁸ process shown in Fig. 6. The injected current I_G is determined by the multiplication M_1 , creating holes that ionize in the substrate with multiplication M_2 , and then reach the floating gate with probability T . For the CISEI process, the programming voltage can be as low as 2.5 V. Figure 6 also shows a structure with a p-Halo to increase the generation of channel-induced substrate electrons.

4. Flash memory structures

Figure 7 shows the world NVSM market since 1980. Notice that the market share of flash memory has increased rapidly since 1990. At present, it has eclipsed both EPROM and EEPROM to become the dominant NVSM with a market share near 70%.

The key reason for this dominance is its small cell size. Figure 8 shows a comparison of cell sizes of various memory devices versus minimum feature length.⁹ SRAM with six transistors (or 4 transistors and 2 load resistors) has, of course, the largest cell size. EEPROM with a selection transistor is also quite large. DRAM is approximately 50% larger than the flash cell because of its storage capacitor. By using a shallow-trench isolation (STI) approach, a flash memory can have a very small cell size.¹⁰ For example, at a 0.25 μm design rule, the cell area is

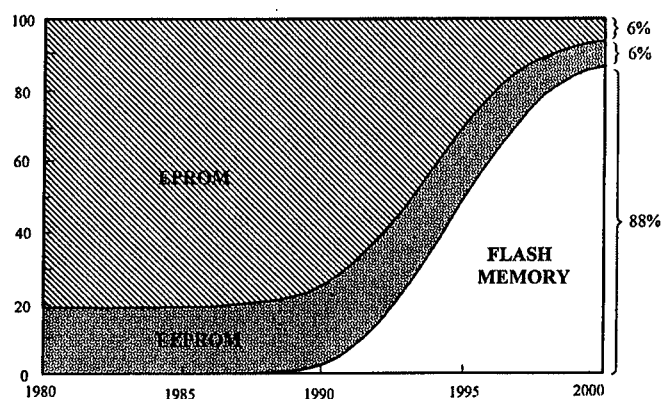


Figure 7. World NVSM market. By the year 2000, flash memory will have nearly 90% of the market share.

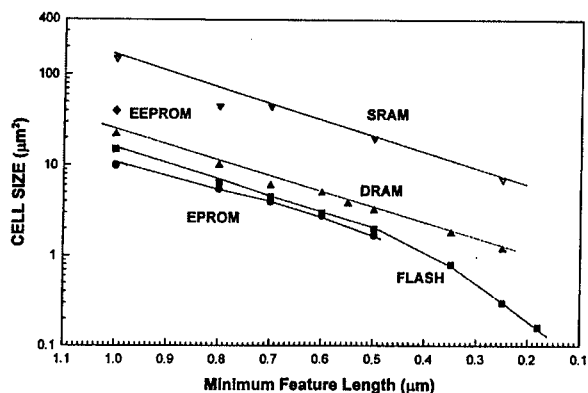


Figure 8. Memory cell size versus minimum feature length for various memory devices.⁹

$0.31 \mu\text{m}^2$, and at a $0.18 \mu\text{m}$ design rule, the cell size is only $0.16 \mu\text{m}^2$. A process flow of the STI-cell is shown in Fig. 9.

Excellent endurance characteristics have been obtained for the STI cell with 9 nm oxide between the floating gate and the channel. The upper threshold voltage remains at 1.5V and shows no change at all after 10^6 program/erase cycles, while the lower threshold voltage varies from -2.0 V to -1.6 V . This memory is expected to maintain an acceptable threshold voltage window of 2V at least to 10^8 program/erase cycles.

Many different cell designs have been used for flash memories.¹¹ These designs can be divided according to data-access and data-write organization. Arrays that have random access and random program (parallel) are consistent with

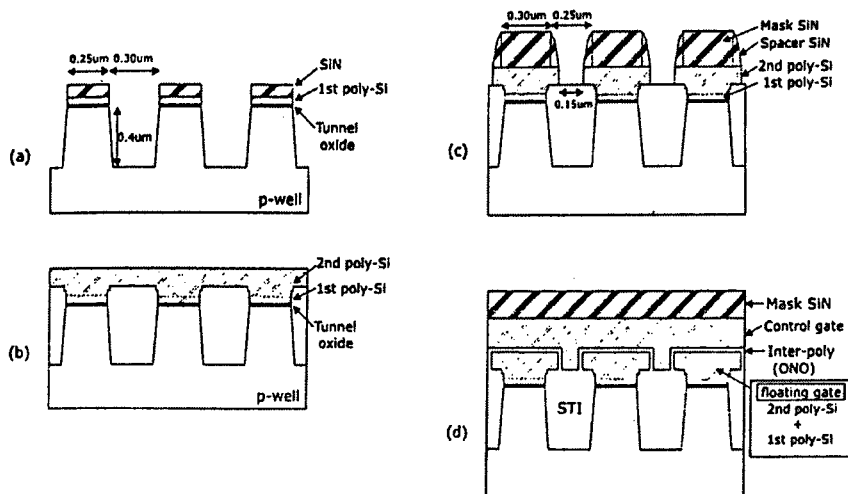


Figure 9. Process flow of the shallow-trench isolation cell. This cell offers very small cell area.¹⁰

embedded applications, while page read and page program (serial) are consistent with mass-storage applications. For programming operations, the different designs use either hot-electron injection or Fowler-Nordheim tunneling. For erasing, Fowler-Nordheim tunneling is used for all designs. In the parallel architectures, we have the asymmetrical contactless transistor, divided bit-line NOR, and common ground standard NOR. In the serial architectures, we have the high-capacitance-coupling-ratio cell, triple-poly virtual-ground cell, and split-gate cell.

5. Single-electron memory cell

The single-electron memory cell (SEMC) is actually a limiting case of the floating-gate structure. By reducing the length of the floating gate, M(1), in Fig. 1, to ultra small dimensions, say 10 nm, we obtain the SEMC. A cross-sectional view of a SEMC is shown in Fig. 10(a). The storage dot corresponds to M(1) in Fig. 1. It is located between the control gate and the channel. Because of its small size the capacitance is also very small (around 10^{-18} F).

The band diagrams of a SEMC are shown in Fig. 10(b). The quantum well can only accommodate one electron.¹² When an electron tunnels into the quantum well, the potential on the left side is reduced and the transfer of another electron is blocked — a result of the "Coulomb blockade". Note that these band diagrams are similar to those shown in Fig. 2. The SEMC is an ultimate floating-gate memory cell, since we need at least one electron for information storage. SEMC operation at room temperature was first demonstrated by Yano and co-workers¹² in 1994 using a floating gate with a gate length of 10 nm. Recently a 128 megabit memory using a vertically unified cell structure of SEMCs has been reported.¹³

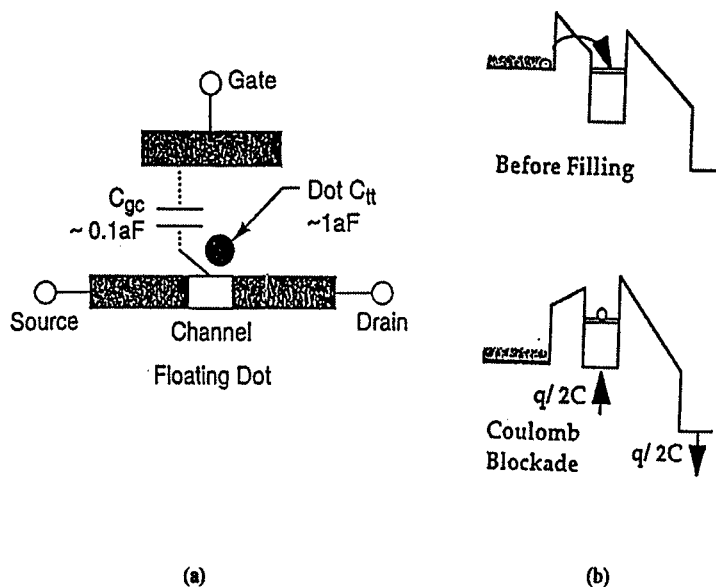


Figure 10. The single-electron memory cell. (a) Cross-sectional view of the device. The floating gate is reduced to about 10nm. (b) Due to the Coulomb Blockade, the information is stored in the form of a single electron.¹²

6. Embedded memory applications

NVSM can be integrated with logic systems (e.g., microprocessors) to form systems on a chip. Such IC chips can have a wide range of applications, including:¹⁴

- Automation: laser printer, ink jet printer, hard disk drive, copier.
- Automobile: powertrain control, automatic braking systems (ABS), instrumentation, navigation.
- Communication: cellular phone, cordless phone, paging, cable TV-set top box.
- Consumer: camera, camcorder, audio recorder, smart card.
- Industry: servo control, motor control, bar code reader.

An example of the embedded memory system is the cellular phone. A block diagram for such system is shown in Fig. 11. A flash memory is embedded in the microcontroller unit along with a microprocessor, a RAM, and a timer.¹⁴ Another example is the single-chip multi-media personal computer that has a flash memory BIOS (basic input and output system) and a flash disk controller.¹⁵

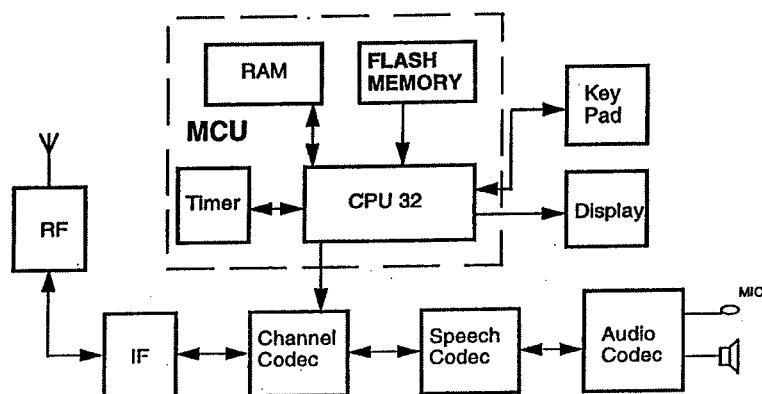


Figure 11. Schematic block diagram of a cellular phone with flash memory.¹⁴

7. Mass storage applications

The hierarchy of memories in a computer system has five levels: the register in the central processing unit (CPU), cache memory, main memory, hard and floppy disks, and magnetic tape¹⁶. The register in the CPU, cache memory, and main memory are generally volatile memories. On the other hand, the hard disk, floppy disk, and magnetic tape are nonvolatile memories. As we move up the hierarchy (i.e., from magnetic tape to the register in CPU), the access time becomes faster; and as we move down the hierarchy, the memory capacity increases. As a result, the market for hard and floppy disks is about five times larger than that for DRAM.

Feature \ Memory	Nonvolatile Memory (Flash Solid State Disk)	Magnetic Disk (2.5" HDD)	Ratio HDD/FSSD
Access Time (ms)	0.2	14	70
Power Consumption (Watt-hours /hour)	0.002	0.4	200
Weight (g/Mbytes)	0.15	4	26
Volume (cm ³ /Mbytes)	0.1	2.9	29
(Operating Shock) ⁻¹ (1/Gs)	0.001	0.1	100
Price (\$/Mbytes)	~1	~0.2	0.2

Table 2. NVSM versus hard disk for portable-computing mass storage.¹⁷

Table 2 shows a comparison of NVSM with hard disk for portable-computing storage.¹⁷ We note that the flash solid-state disk is 70 times faster in access time, 200 times lower in power consumption, 26 times lighter in weight, 29 times smaller in size, and 100 times better in ruggedness than a hard disk. At the present time, the only advantage of a hard disk is its price. However, the price per Mb of NVSM is falling faster than that of a hard disk. It is, therefore, expected that in a few years NVSM will replace most of the hard and floppy disks.

8. Technology drivers

The electronic industry is the largest industry in the world with global sales of over 1 trillion U.S. dollars. The foundation of the electronic industry is the semiconductor industry. Figure 12 shows the sales volumes of these two industries in the past ten years and projects the sales to year 2010. If the current trends continue, the sales volumes of the electronic industry and the semiconductor industry will reach \$3 trillion and \$600 billion, respectively, in the year 2010.

Also shown in Fig. 12 are the sales volumes for DRAM, SRAM and NVSM. The global sales of NVSM are about \$6 billion with an impact on the system level well over \$100 billion. NVSM has surpassed SRAM in 1996. If it maintains a 35% to 45% annual growth rate, NVSM will surpass DRAM in about 5 years to become the dominant memory for the electronic industry.

The growth curves for different technology drivers are shown¹⁶ in Fig. 13. At the beginning of the modern electronic era (1950 to 1970), the bipolar transistor was the technology driver. From 1970 to 1990, the DRAM and microprocessor were the technology drivers due to the rapid growth of personal computers and

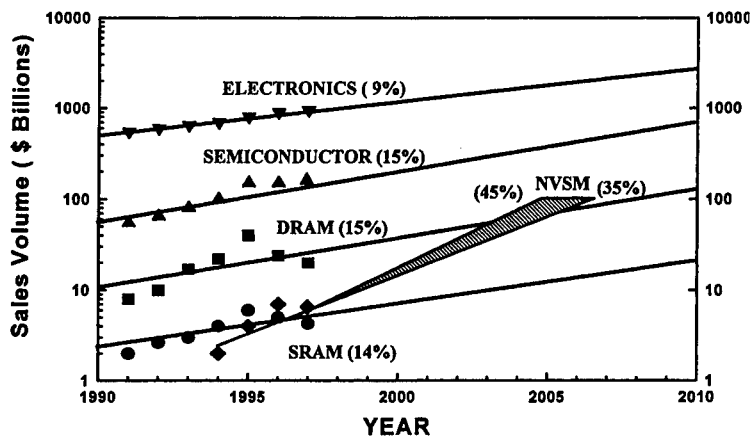


Figure 12. Sales volume of electronic industry, semiconductor industry, DRAM, SRAM and NVSM. Sales of the NVSM overtook SRAM in 1996 and will surpass DRAM in 5 to 10 years.

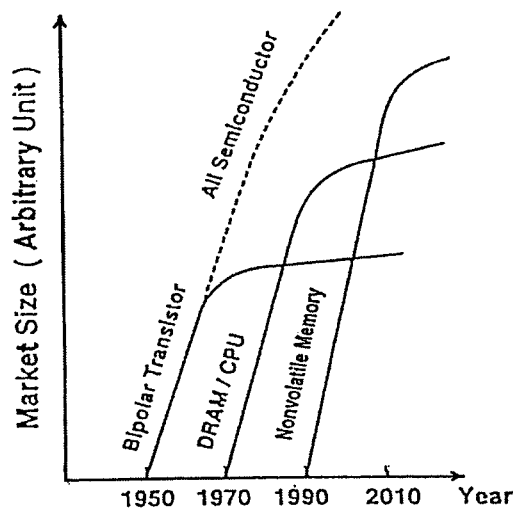


Figure 13. Since 1990, NVSM has been the technology driver of the semiconductor industry.¹⁶

advanced electronic systems. Since 1990, NVSM has been the technology driver. This is because of the rapid growth of portable electronic systems such as the cellular phone, notebook computer, smart card, etc, which need memories having the most desirable attributes.

9. Conclusions

We have presented a brief review of the historical developments of nonvolatile semiconductor memory (NVSM) and projected its trends to year 2010. In the past 30 years, NVSM has emerged from a simple floating-gate concept to a multi-billion-dollar industry. A key member of NVSM family is the flash memory, which has captured about 70% of the NVSM market. The ultimate flash memory is the single-electron memory cell, which may serve as a building block for 1 Tb (10^{12} bit) nonvolatile memory.

Because of its attributes of high density, low-power consumption, nonvolatility, and electrical rewritability, NVSM has become the dominant memory for portable electronic systems such as cellular phone, notebook computer, smart IC card, etc. NVSM has surpassed SRAM in sales volume. It is now poised to eclipse DRAM's market share and to replace hard and floppy disks in the near future.

References

1. 1997 *Electronic Market Data Book*, Electronic Industries Association, Washington, D. C., 1997.
2. A. K. Sharma, *Semiconductor Memories — Technology, Testing, and Reliability*, Piscataway: IEEE Press, 1997.
3. W. D. Brown and J. E. Brewer, eds., *Nonvolatile Semiconductor Memory Technology*, New York: IEEE Press, 1998.
4. D. Kahng and S. M. Sze, "A floating gate and its application to memory devices," *Bell Syst. Tech. J.* **46**, 1288 (1967).
5. D. Frohman-Bentchkowsky, "Memory behavior in a floating-gate avalanche-injection MOS (FAMOS) structure," *Appl. Phys. Lett.* **18**, 332 (1971).
6. H. Iizuka, F. Masuoko, T. Sato, M. Ishikawa, "Electrically alterable avalanche injection type MOS read-only memory with stacked gate structure," *IEEE Trans. Electron Dev.* **ED-23**, 379 (1976).
7. F. Masuoko, M. Asano, H. Iwahashi, T. Komuro, and S. Tanaka, "A new flash E²PROM cell using triple polysilicon technology," *IEDM Tech. Dig.*, 464 (1984).
8. J. Bude, A. Frommer, M. Pinto, and G. Weber, "EEPROM/Flash sub 3.0 V drain-source bias hot-carrier writing," *IEDM Tech. Dig.*, 990 (1995).
9. S. Lai, "The outlook of flash EPROM technology," *Proc. VLSI Technology, Systems, Applications* (May 12, 1993), p. 147.
10. K. Shimizu, K. Narita, H. Watanabe, E. Kamiya, Y. Takeuchi, T. Yaegashi, S. Aritome, and T. Watanabe, "A novel high-density 5F² NAND STI cell technology suitable for 256 Mbit and 1 Gbit flash memories," *IEDM Tech. Dig.*, 271 (1997).
11. P. Pavan, R. Bez, P. Olive, and E. Zanoni, "Flash memory cells — an overview," *Proc. IEEE* **85**, 1248 (1997).
12. K. Yano, T. Ishii, T. Hashimoto, T. Kabayashi, F. Murai, and K. Seki, "Room temperature single-electron memory," *IEEE Trans. Electron Dev.* **41**, 1628 (1994).
13. K. Yano, T. Ishii, T. Mine, F. Murai, T. Kure, and K. Seki, "128 Mb early prototype for gigascale single-electron memories," *ISSCC Tech. Dig.* **41**, 344 (1998).
14. C. Kuo, "Embedded flash memory, application, technology, and design," IEDM Short Course on NVRAM Technology and Applications (1995).
15. H. Sasaki, "Multimedia future and impact for semiconductor technology," *IEDM Tech. Dig.*, 3 (1997).
16. F. Masuoka, "Flash memory technology," *Int. Electron. Dev. Mater. Symp.*, 83 (1996).
17. A. G. Barre, "Flash memory, magnetic disk replacement?" *IEEE Trans. Magn.* **29**, 4104 (1993).

Double-Junction Gated Single-Electron Transistor EEPROM Cell

M. Y. Jeong, Y. H. Jeong, and D. M. Kim

Department of Electronic and Electrical Engineering, Pohang University of Science and Technology (POSTECH), San-31, Hyoja-dong, Nam-ku, Pohang, Kyungbuk, 790-784, South Korea

1. Introduction

The memory embedded logic is an essential component of the system-on-a-chip and a key requirement of the memory technology is its integrability. Also, with the portable electronic system quickly becoming prevalent, nonvolatility constitutes another attractive feature of the memory. The flash EEPROM technology satisfies both of these requirements, together with high density, low power, and repeated fast programming with appropriate endurance.

At present, programming a flash EEPROM cell requires about 5×10^4 electrons to be stored on the storage node. Theoretically, single electronics enables memory operation with the use of a single electron. Due to this low power consumption and high density integration, the single electron transistor (SET) memory is viewed as one of the promising applications of single electronics. The purpose of this paper is to present a simple design of a SET memory cell and discuss its operational principles.

The designed cell consists of an SET, in which a double tunnel junction trap is incorporated in the gate structure to provide the storage node for electrons. The device operation is based on the controlled shift of the threshold voltage of the SET, depending on the presence of excess electrons on the storage node. The characteristics of the SET memory cell, as simulated by the Monte Carlo method, are presented and discussed in analogy with flash EEPROM cells.

2. Operational principle

As mentioned, the operation of the SET memory cell is based on the controlled shift of the SET transfer curve along the gate voltage axis. Its operation can conveniently be discussed by pointing out the similarities and differences between SET and flash memory cells. For this purpose the operation of the flash EEPROM cell is briefly reviewed.

Since the nonvolatile semiconductor memory device was first proposed by Kahng and Sze,¹ a number of other similar devices has been introduced. Figure 1 shows the cross-section of the simplest type, namely the stacked-gate NOR-type

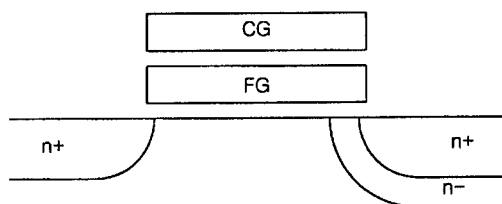


Figure 1. Schematic cross-section of a stacked-gate flash EEPROM cell.

flash EEPROM.² The structure is similar to a MOSFET except that a floating gate is inserted between the channel and the control gate.

The programming is done via the generation of hot electrons by the channel pinch-off induced high field near the drain and injection of some of these electrons onto the floating gate. If the resulting excess electronic charge is $-Q$, the threshold voltage at the control gate is shifted by

$$\Delta V_{TH} = \frac{Q}{C_c} \quad (1)$$

Here C_c is the capacitance between the control and floating gates. The quantity, $C_c \Delta V_{TH}$, represents the charge that should be induced at the control gate to compensate for $-Q$. The programming thus consists of inducing a desired ΔV_{TH} . Erasing is done by releasing the stored electrons via Fowler-Nordheim tunneling.

The key features of the memory cell are the speed and efficiency of programming. These are determined by two factors: the generation and injection of hot electrons onto the floating gate. These two processes are statistical in nature and vary widely depending on the device structure, as detailed elsewhere.³ For the stacked gate cell, for example, typically one out of 10^7 electrons streaming down the channel from the source is used for programming. This low efficiency necessitates the use of large power or voltage.

Figure 2 shows the schematic of the SET memory cell proposed here. Tunnel junctions (TJ) 1 and 2 with capacitances C_G and C_B constitute a SET, while TJs

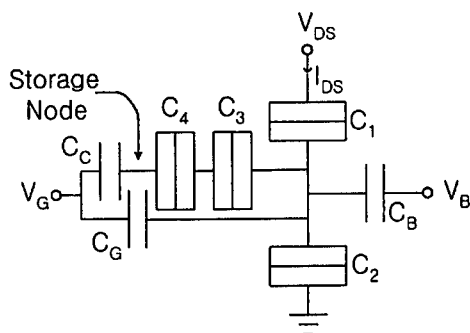


Figure 2. Schematic of single-electron transistor memory cell.

3 and 4 with C_C form the double tunnel junction (DTJ) electron trap. For the trap one can also use a multiple-tunnel junction (MTJ)⁴ or a dual-junction-array (exciton trap).⁵ The DTJ and SET play the roles of the floating gate and the channel underneath, respectively. The bistability inherent in the DTJ electron trap together with the usual periodic oscillation of the SET transfer characteristics provide the basis for the memory operation.

Figure 3 illustrates the transfer characteristics of both flash and SET memory cells. In the former V_{TH} represents the channel inversion, giving rise to an "on" current under bias. In the SET V_{TH} is defined as the gate voltage at which the channel Coulomb blockade is lifted, resulting in current conduction.

Consider the case where TJs 3 and 4 are in the Coulomb blockade state with zero excess charge on all islands. Then the memory cell reduces to a SET with total gate capacitance, $C_{GT} = C_G + C_C/C_3/C_4$, with $//$ denoting the series connection.

As V_G increases, TJ 2 is released from the blockade state, and the SET begins to conduct. The threshold voltage for conduction is given by

$$V_{TH} = \frac{1}{C_{GT}} \left[\frac{e}{2} - C_1 V_{DS} - C_B V_B \right]. \quad (2)$$

With further increase of V_G , the voltage across the double junction trap should also increase and at a specific voltage, the Coulomb blockade of TJ 3 is released. In this case an electron tunnels through the junctions 3 and 4 in succession to be trapped in the storage node. This occurs at V_{WT} given by

$$V_{WT} = \frac{1}{C_\alpha} \left[-\frac{e}{2} + \frac{e}{2} \frac{C_\beta}{C_4 // C_C} + \frac{C_\beta}{C_C} n_{sn} e + C_B V_B + C_1 V_{DS} \right]. \quad (3)$$

Here, n_{sn} is the number of electrons in the storage node, $C_\alpha = C_1 + C_2 + C_B$, and $C_\beta = C_\alpha + C_G$.

In the SET memory cell, the programming is performed by driving TJs 3 and 4 into conduction via an appropriate choice of gate voltage, in contrast with the

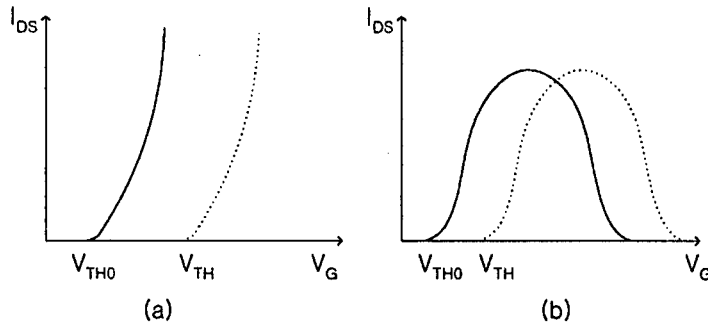


Figure 3. Drain current versus gate voltage in (a) flash and (b) SET memory cells with (dotted line) and without (solid line) the charge in the storage node.

case of a NOR-type flash EEPROM cell, in which hot electrons are to be generated first and only a minor fraction of these electrons are statistically injected into the floating gate for programming. Hence in the SET memory cell the programming is more controllable and should have much higher efficiency. Also, the oscillatory transfer $I(V)$ curve of the SET shows that the current ratio of the programmed/unprogrammed cell also becomes an oscillatory function of the threshold voltage shift. Thus, one period of oscillation is sufficient as a threshold voltage shift to get a large current ratio at the read gate voltage.

The excess electrons on the storage node induce a shift of the SET transfer curve by an amount

$$\Delta V_{TH} = \frac{1}{C_{GT}} \frac{C_F}{C_F + C_C} n_{sn} e. \quad (4)$$

Here $C_F = C_3/C_4$ the capacitance between the storage node and the center island of the SET. When $C_G = 0$, Eq. (4) is reduced to Eq. (1), i.e. $en_{sn} = C_C \Delta V_{TH}$, again suggesting that ΔV_{TH} comes about to induce gate charge to compensate for the trapped electronic charge, $-en_{sn}$. The presence of C_G obviously renders the required gate charge larger than en_{sn} , as it should.

3. The simulation

The characteristics of the SET memory cell have been simulated using the Monte Carlo method.⁶ The values of capacitance used are scaled with respect to the tunnel capacitance C_2 as: $C_1 = C_2$, $C_3 = C_4 = 0.5C_2$, $C_C = 6C_2$, $C_G = 0.2C_2$, and $C_B = 0.4C_2$. All tunnel resistances are set to a constant value R , which is taken much larger than the resistance quantum, R_Q . The operating temperature of the cell is taken as $T = 1.6 \times 10^{-3} e^2/k_B C_2$ where k_B is the Boltzmann constant.

Figure 4 shows the transfer curves of the SET memory cell when there are no electrons stored on the storage node (curve A) and when three electrons are stored on the storage node (curve B). Here the supply drain voltage was chosen to be $V_{DS} = 0.15e/C_2$ and $V_B = 0$. As shown in Fig. 4, the electron tunneling onto the storage node shifts the transfer curve to the right — see Eqs. (2) and (3). Three electrons are transferred to the storage node at the bias condition chosen, as exhibited by the three sawtooth-like steps. With C_1 of 0.1 aF, the threshold voltage shift ΔV_{TH} amounts to about 0.435 V, which can be readily sensed.

In the flash EEPROM cell the number of electrons injected onto the floating gate increases monotonically with programming time and saturates at a level mainly dictated by the control gate voltage. In the SET memory cell, however, the number of electrons stored is self-limited by the Coulomb blockade.⁷ That is, the maximum number of electrons that can be stored is determined by the cell parameters. With the device parameters used, four electrons can be stored in the storage node in the simulation and the maximum threshold voltage shift that can be obtained is about $0.364e/C_2$.

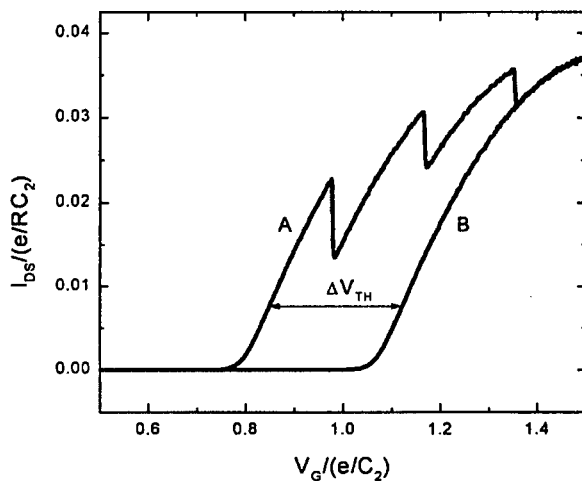


Figure 4. $I(V)$ curves of a SET memory cell when there is no electrons stored in the storage node (curve A) and when three electrons are stored in the storage node (curve B).

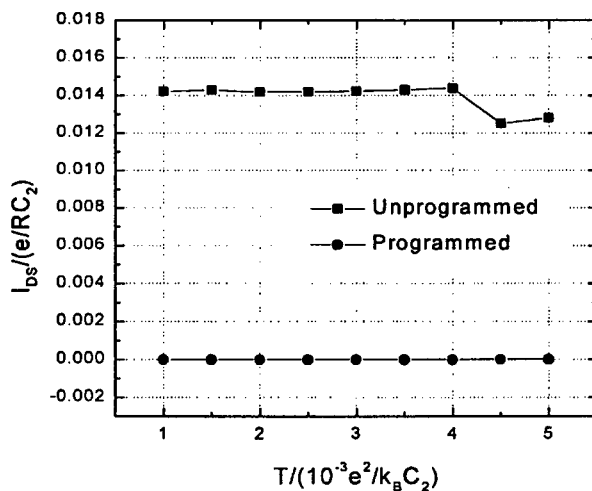


Figure 5. Drain current of programmed/unprogrammed cell vs. temperature at read bias condition.

In Fig. 5, the calculated drain current of the programmed/unprogrammed cell is presented at the read bias condition of $V_{DS} = 0.15e/C_2$, $V_G = 0.9e/C_2$. The perfect condition in which $n_{sn} = 0$ and $n_{sn} = 3$ for unprogrammed and programmed states, respectively, persists up to about $T = 4.25 \times 10^{-3} e^2/k_B C_2$. With C_2 of 0.1 aF,

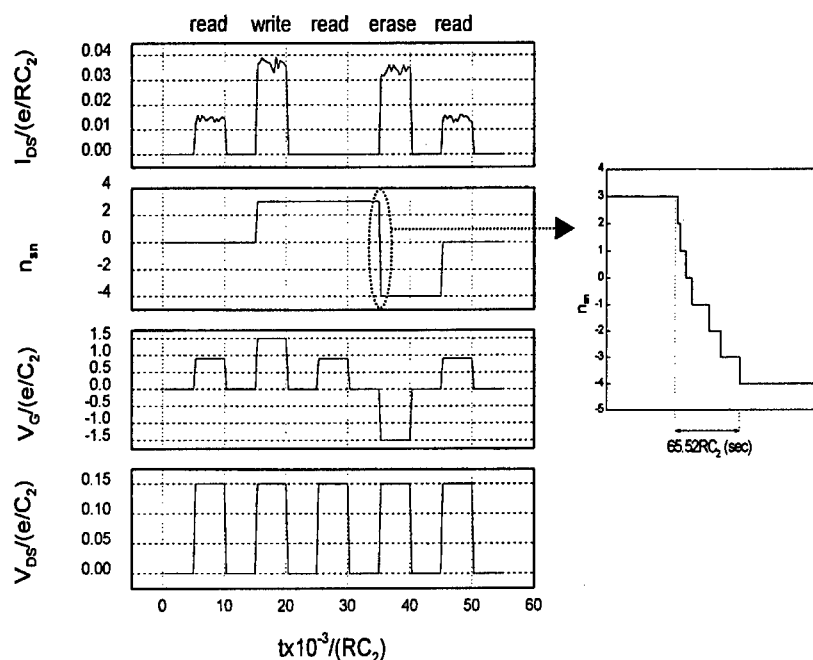


Figure 6. Drain current, number of trapped electrons and read/write/erase voltages vs. time. The inset shows the releasing of electrons from the trap site during erase.

the value of T is about 79 K. At higher temperatures, n_{sn} in the unprogrammed cell increases due to thermal activation, degrading thereby the read current. With appropriate relaxation of the read conditions, the maximum operating temperature can be raised further.

Figure 6 shows the program/erase/read characteristics of the cell versus time. The read/write gate voltages are set as $0.9e/C_2$ and $1.5e/C_2$ respectively, with the common drain voltage of $0.15e/C_2$. The average current in the unprogrammed cell at the read gate voltage is about $14.2 \times 10^{-3} e/RC_2$, i.e. 22.8 nA with typical values of $C_2 = 0.1$ aF and $R = 1$ M Ω . For the same values of these parameters the voltage pulses have time duration of 1 ns. For the programming conditions chosen, three excess electrons are transferred to the storage node in a time span of approximately 3 ps and the reading after programming indicates that I_{DS} is practically zero, as it should be. For the given erase bias conditions the three excess electrons as well as four additional electrons are released in about 7 ps to ensure complete erasure. The net number of electrons in the storage node recovers the initial value of zero during the read after erasure. When the erasure is immediately accompanied by programming, seven electrons are transferred to the storage node, so that the same three excess electrons end up in storage. It is clear from the simulation that the SET memory cell is capable of fast programming and erasing with a tight distribution of V_{TH} shift and a large read margin.

4. Conclusions

A simple design of a nonvolatile SET memory cell has been presented and discussed using Monte Carlo simulations. The operation of the device has been discussed in analogy with and in comparison with the conventional flash EEPROM cell. In the latter cell the programming is performed typically in 10 μ s by using the supply (drain) current of a few hundred μ A. Also, approximately 5×10^4 electrons are stored in the floating gate to induce a V_{TH} shift of about 5 V.

In the SET memory cell, our calculations show that the programming can be achieved in approximately 10 ps using a supply current of about 50 nA. Furthermore, fewer than 10 electrons are needed on the storage node to induce a detectable V_{TH} shift. Additionally, both programming and erasing are fully controlled by the Coulomb blockade. In spite of these attractive features, the SET memory approach is marred by the difficulty of processing nanostructures and low temperature operating conditions.

It may therefore be worthwhile to investigate the possibility of relaxing the stringent geometry of the SET memory cell without losing too much of its advantages. For example, with the use of a large control gate capacitance, it is possible to increase the maximum allowed storage node electron number. This, in turn, reduces the degradation of the read-out current arising from the fluctuation of the n_m . Thus the error rate and the maximum operating temperature could be improved.

Multiple tunnel junctions could also mitigate various problems. Macroscopic quantum tunneling can be reduced due to the fact that the co-tunneling error rate is inversely proportional to the product of the resistances of the tunnel junctions composing the MTJ. The tunnel junctions can be made large in an MTJ array, since the serial connection of tunnel junctions results in a small effective capacitance and a large resistance. These considerations provide some room for relaxing the size requirement of the SET memory cell.

Additionally, a tunnel junction with a given capacitance can be realized with a larger capacitor with thicker oxide, although the thicker oxide will reduce the speed. However, in the final analysis the SET memory cell is based on the Coulomb blockade and the real challenge consists of devising process techniques by which to implement the room temperature Coulomb blockade. Once this process hurdle is overcome the advantages of the SET memory cell could be fully fused into the nonvolatile semiconductor memory technology.

References

1. D. Kahng and S. M. Sze, "Floating gate and its application to memory devices," *Bell Syst. Tech. J.* **46**, 1283 (1967).
2. S. Mukherjee and T. Chang, "Single transistor electrically programmable memory device and method," U.S. patent 4,698,787.

3. D. M. Kim, M. K. Cho, and W. H. Kwon, "Stacked gate mid-flash EEPROM cell. Part I: programming speed and efficiency versus device structure," *IEEE Trans. Electron Dev.* **45**, 1696 (1998).
4. K. Nakazato and H. Ahmed, "The multiple tunnel junction and its application to single-electron memory and logic circuits," *Jpn. J Appl. Phys.* **34**, 700 (1995).
5. S. Amakawa, M. Fujishima and K. Hoh, "Dual-junction-array single electron trap," *Abstracts Electrochem. Soc.* **96-2**, 572 (1996).
6. M. Y. Jeong, Y. H. Jeong, S. W. Hwang, and D. M. Kim, "Performance of single-electron transistor logic composed of multi-gate single-electron transistors," *Jpn. J. Appl. Phys.* **34**, 6706 (1997).
7. L. G. Guo, E. Leobandung, and S. Y. Chou, "A room temperature silicon single-electron metal-oxide-semiconductor memory with nanoscale floating-gate and ultranarrow channel," *Appl. Phys. Lett.* **70**, 1742 (1997).

Single-Electron Memories with Terabit Capacity and Beyond

C. Wasshuber

Texas Instruments Inc., 13570 North Central Expressway, Dallas, Texas 75243, USA

H. Kosina and S. Selberherr

Institute for Microelectronics, TU Wien, Gußhausstraße 27-29/E360, A-1040 Wien, Austria

1. Introduction

The need for more information storage capacity in the near future is evident. The lasting trend to mobile electronics (palm computers, cellular phones, video cameras, still cameras, etc.) demands high density and low power storage devices. Take for example a high resolution color still image ($1024 \times 768 \times 24$ bit). Without compression one requires almost 20 Mbit storage capacity for a single image. This relates to about 2 Gbit for a still camera capable of storing 100 images, and to almost 4 Tbit for a 2 hour high resolution video (30 frames/second). Certainly intelligent compression techniques can reduce these numbers considerably. We mention them here to give the reader a feeling for the magnitudes involved. Or look for example at software packages. They need frequently more than 1 Gbit storage capacity, and the exponential growth continues. In order to meet these demands and to provide product engineers with the desired memory components, a fast, high density, low power and possibly nonvolatile memory technology is required. A solid state solution has a clear speed (access time) and miniaturization advantage over magnetic or optical solutions. However, conventional solid state memory has a storage size, power consumption, and cost/bit disadvantage to other technologies. Single-electron technology provides great prospects to meet these challenges. Thus, we will present a single-electron memory cell, our T-memory cell, which can be integrated with today's process technology to produce memory chips of 1 Tbit storage capacity and beyond. At the same time power consumption is reduced by orders of magnitude and an access time of about 1 ns seems possible.

Although single-electron technology seems today to be more appropriate for memories, a new promising scheme for single-electron logic has been proposed.¹ Especially its reduced sensitivity to random background charge gives hope for large scale integrated single-electron logic devices.

In Section 2 we give a brief introduction to single-electronics. In Section 3 we present our analysis tool, SIMON, a single-electron device and circuit

simulator. In Section 4 we discuss the operation, production, advantages and disadvantages of the T-memory cell.

2. Short introduction to single-electronics

Quantization of charge in metallic or semiconductor islands (quantum dots, granules) is usually not directly noticeable. However, when the size of such islands is in the nanometer regime, that is when the total capacitance becomes very small and the charging energy is larger than the thermal energy, then the change in free energy associated with the addition or subtraction of a single electron from an island, or a quantum dot, becomes significant.

New phenomena appear, such as the Coulomb blockade, which is a suppression of current flow at low voltage bias, and Coulomb oscillations, a time or space correlated transfer of electrons through tunnel junctions. With these new quantum effects it is possible to control the movement and position of single electrons. Beside the desired characteristics of controlled transfer of single electrons, undesirable effects arise, too. Among these, co-tunneling and the sensitivity to uncontrollable impurities and stray charges are the most critical ones. Co-tunneling, a simultaneous tunneling of two or more electrons in different tunnel junctions, provides a path for electrons to escape the Coulomb blockade and thus "weakens" the Coulomb blockade. Co-tunneling often plays an important role in the lifetime of stable electron states, for example stored electrons representing a bit of information, and in leakage currents. The sensitivity to uncontrollable impurities and stray charges, which is referred to as sensitivity to random background charges, introduces another serious handicap. Random background charge directly "attacks" the size of the Coulomb blockade and can in certain cases eliminate the Coulomb blockade completely. Whereas the amount of co-tunneling can be engineered, by introducing less transparent tunnel junctions or introducing more tunnel junctions in series, random background charge is much more difficult to deal with. In the case of single-electron memories practical ideas to eliminate the random background charge problem exist, as will be explained later, but for single-electron logic no real practical approach has been developed.

A general introduction to the field of single-electronics can be found for example in Ref. 2 and a very detailed and comprehensive one in Ref. 3.

3. Analysis tool — SIMON

To conduct our analyses we developed a simulation tool. SIMON is a single-electron device and circuit simulator based on a Monte Carlo method, where the free energy before and after any particular tunnel event determines the probability for tunneling. Among all possible events one is chosen as the winner according to the computed probability distribution and the state of the circuit is updated. By simulating many tunnel events the macroscopic device characteristics are obtained. In the same manner co-tunneling can be accounted for, by allowing two or more

tunnel events to happen at the same time. One important condition underlying the successful operation of single-electron devices and also underlying this simulation method is that electrons have to be well localized on the islands. This condition means that all tunnel resistances have to fulfill

$$R_T > \frac{h}{e^2} \equiv 25813 \, \Omega \quad (1)$$

$$\Gamma(\Delta F) = \frac{1}{e^2 R_T} \left(\frac{-\Delta F}{1 - e^{\frac{\Delta F}{k_B T}}} \right) \quad (2)$$

$$\begin{aligned} \Gamma^{(N)} = & \frac{2\pi}{\hbar} \left(\prod_{i=1}^N \frac{\hbar}{2\pi e^2 R_{T_i}} \right) \left[\sum_{\text{perm}(k_1, \dots, k_N)} \left(\prod_{k=1}^{N-1} \frac{1}{\Delta F_k - \frac{k}{N} \Delta F_N} \right) \right]^2 \times \\ & \times \frac{\Delta F_N}{(2N-1)! \left(e^{\frac{\Delta F_N}{k_B T}} - 1 \right)} \prod_{i=1}^{N-1} \left((2\pi k_B T_i) + \Delta F_N^2 \right) \end{aligned} \quad (3)$$

$$\tau = -\frac{\ln(r)}{\Gamma} \quad (4)$$

where ΔF is the change in free energy caused by a single tunnel event, Γ is the tunnel rate for normal tunnel events, $\Gamma^{(N)}$ is the tunnel rate for N^{th} -order co-tunneling, τ is the time duration between two consecutive tunnel events, and r denotes a uniformly distributed random number in the interval $[0, 1]$. More information on the simulation of single-electron devices and circuits can be found in Ref. 4.

SIMON features a graphical circuit editor that is embedded in a graphical user interface (Fig. 1). This interface allows easy use and quick circuit design. An interactive simulation mode is provided, where electrons can be forced to tunnel through particular junctions. Node charges, node voltages, tunnel rates and energy differences can be studied interactively. This mode allows a very detailed analysis of single-electron circuits. SIMON is publicly available. For further information see Refs. 5-7.

4. T-memory

High-density memories are a very suitable application for single-electron devices. The storage of binary information using Coulomb-blockade phenomena has been demonstrated in theory^{3,8} and in practice.⁹⁻¹² Just recently, at the 1998 International Solid State Circuits Conference, Yano *et al.* presented a 128 Mbit early prototype for gigascale single-electron memory.¹³ This chip is so far the biggest single-electron memory fabricated. However, several issues were identified

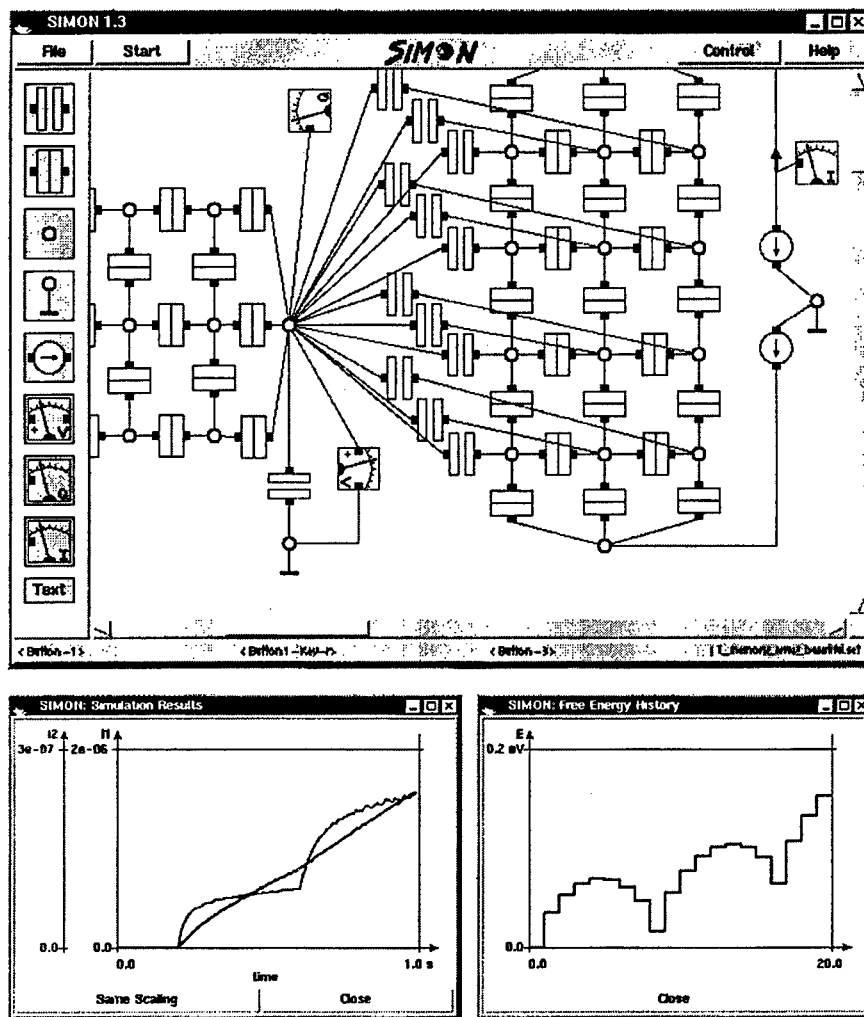


Figure 1. Screen shots of SIMON. The top picture shows the graphical circuit editor. In the lower left picture the I-V characteristics of a symmetric and asymmetric single-electron transistor are shown. In the lower right picture the change in system energy for the charging of a single-electron trap is shown.

that make a large scale integration of single-electron memory cells with today's process technology very difficult. The most notable of these problems are the achievement of room temperature operation, which demands lithographic resolution of less than 10 nm, and the sensitivity to random background charge, which demands perfectly clean production processes. Considering the current ULSI trends, neither of these issues will be resolved in the near future.

Summarizing, the main requirements to make single-electron circuits feasible for industrial application today are:

- room temperature operation → islands smaller than 10 nm
- insensitivity to random background charge
→ absolutely clean production processes
→ Coulomb oscillations instead of Coulomb blockade
- mass production → optical lithography

All of these requirements can be met by using granular films. Their usage eliminates the necessity for nanolithography, since nanoscopic tunnel junctions and granules are formed naturally. Additionally, granular films provide an averaging of the Coulomb blockade, alleviating the background charge problem.

Several production methods for two-dimensional granular films spanning a wide variety of material systems have been developed and many more are under development. Metallic granular films can be produced by condensing Al or Au on a substrate.¹⁴ Several methods exist in which semiconducting granular films are formed.¹⁵⁻¹⁷ Employing molecular beam epitaxy, nanoscopic InGaAs or InP dots or pyramids can be grown.^{18,19} Colloidal deposition can be used to first form granules, which are afterwards deposited on a substrate.^{20,21} This process is possible for metallic and semiconducting materials. Since single-electronics requires only small conducting granules separated by thin nonconducting tunnel barriers, more exotic material systems are equally interesting. For example, polymers of which some forms can conduct current could be used. Or even organic films would be a possibility. This latitude in material systems gives great promise as well as a lot of work for the future of single-electronics.

Several memory designs were analyzed with SIMON.²² In consequence of these investigations, we proposed a new memory design called T-memory. This cell consists of two granular film batches that are arranged in T-form and that are capacitively coupled at their junction. Figure 2 shows two possible realizations.

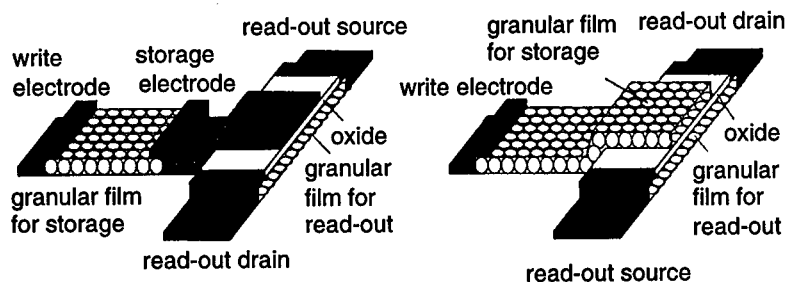


Figure 2. The T-memory consists of two granular film batches that are capacitively coupled. On the left side the capacitive coupling is achieved with a separate electrode. On the right side, an alternative realization, the two granular film batches directly overlap.

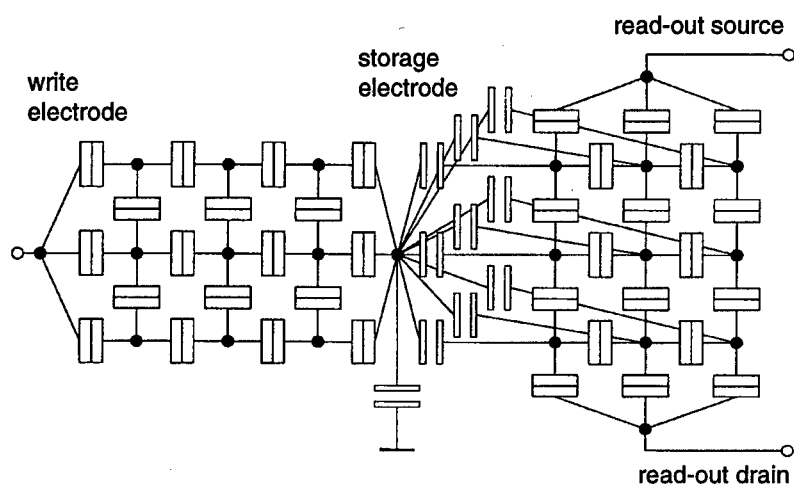


Figure 3. Equivalent circuit of the T-memory cell.

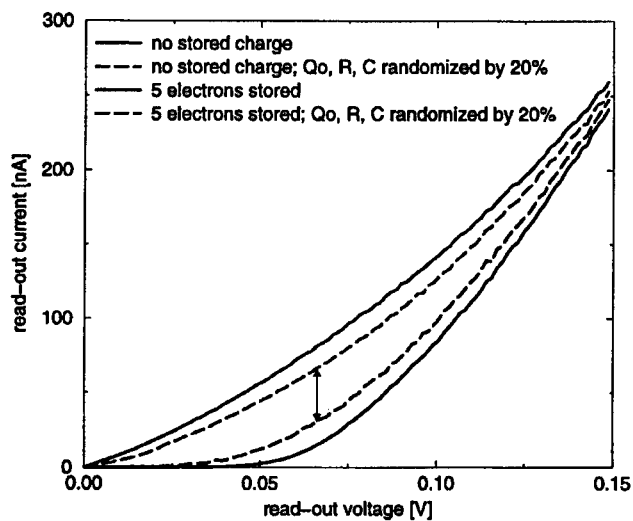


Figure 4. Charge dependence of the $I(V)$ characteristic of the granular film electrometer (transistor).

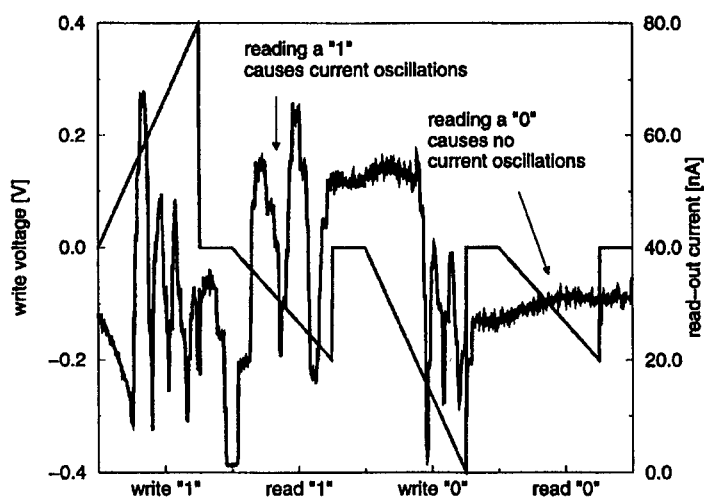


Figure 5. Second read-out mechanism employing Coulomb oscillations. The stored information can be retrieved by discharging the storage electrode.

Figure 3 depicts the equivalent circuit of the T-memory cell. It turns out that using two granular films, one for storing charge and the other to read out the stored information, gives the designer great flexibility in optimizing the cell. This flexibility enables one to tune storage time to achieve, for example, a nonvolatile memory and at the same time to optimize read-out characteristics for reliable information retrieval.

The granular batch representing the cross-bar of the T can be viewed as a multi-island single-electron electrometer (transistor). The other granular batch is the port to charge or discharge the storage electrode that influences the electrometer.

Two different read-out mechanisms are imaginable for the T-memory cell. The first one makes use of the dependence of the static $I(V)$ characteristic of the read-out electrometer on the stored charge. Figure 4 shows the dependence of the $I(V)$ characteristic on the charge stored in the storage electrode.

The other mechanism is a destructive read-out, which was first published by Likharev and Korotkov.²³ One attempts to discharge the storage node. If current oscillations are sensed with the read-out electrometer, the cell was charged (state 1), otherwise it was empty (state 0). Accordingly, the contents of the cell have to be restored. This operation is similar to a refresh cycle in DRAMs. Figure 5 shows the signals when reading a 1 and a 0.

Arranging several T-memory cells in matrix form results in a bit-addressable memory chip. Applying at the same time a positive voltage pulse at a word line and a negative pulse at two adjacent bit lines (or *vice versa*, a negative pulse at the

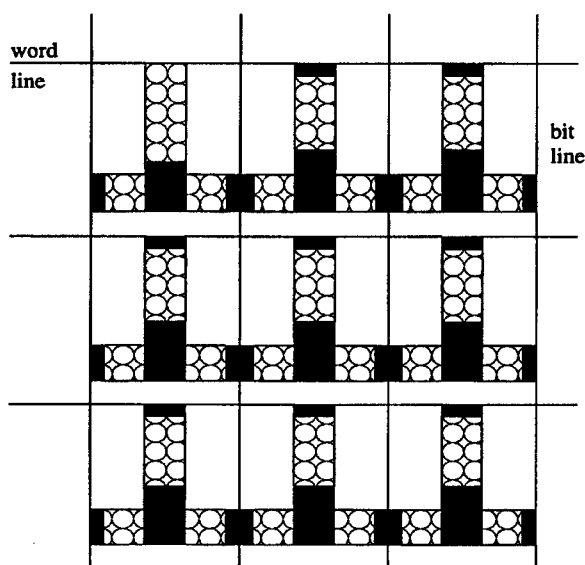


Figure 6. Bit-addressable T-memory. Applying at the same time a positive voltage pulse at a word line and a negative pulse at two adjacent bit lines addresses one cell, which can be written or read.

word line and a positive pulse at the bit lines) addresses one cell, which can be written or read (see Fig. 6). The bit-addressability is only possible if the Coulomb oscillation read-out mechanism is used.

5. Conclusion

Attempting to solve the main challenges, which are room temperature operation, random background charge independence, and industrial mass fabrication with today's process technology, we have proposed the T-memory cell. This memory cell employs two capacitively coupled granular films. The films can be made from various materials ranging from metals and semiconductors to polymers and organic substances. Nanometer granules that make up the granular film provide a Coulomb blockade large enough for room temperature operation. Granules form naturally, thus there is no need for nanolithography. Read-out by Coulomb oscillations and the randomizing statistical properties of granular films provide considerable insensitivity to random background charges. Thus the T-memory cell can be integrated with today's process technology for terabit memory chips.

References

1. T. Oshima and R. A. Kiehl, "Operation of bistable phase-locked single electron tunneling logic elements," *J. Appl. Phys.* **80**, 912 (1996).
2. K. K. Likharev, "Correlated discrete transfer of single electrons in ultrasmall tunnel junctions," *IBM J. Res. Develop.* **32**, 144 (1988).
3. H. Grabert and M. H. Devoret, eds., *Single Charge Tunneling—Coulomb Blockade Phenomena in Nanostructures*, New York and London: Plenum Press and NATO Scientific Affairs Division, 1992.
4. C. Wasshuber, *About Single-Electron Devices and Circuits*, Ph.D. thesis, Technische Universität Wien, 1997.
5. C. Wasshuber, H. Kosina, and S. Selberherr, "SIMON—a simulator for single-electron tunnel devices and circuits," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.* **16**, 937 (1997).
6. C. Wasshuber and H. Kosina, "A single-electron device and circuit simulator," *Superlatt. Microstruct.* **21**, 37 (1997).
7. Information available at <http://www.iue.tuwien.ac.at/>.
8. K. Nakazato, R. J. Blaikie, and H. Ahmed, "Single-electron memory," *J. Appl. Phys.* **75**, 5123 (1994).
9. K. Yano, T. Ishii, T. Hashimoto, *et al.*, "Room-temperature single-electron memory," *IEEE Trans. Computer-Aided Design* **41**, 1628 (1994).
10. N. J. Stone and H. Ahmed, "Silicon single-electron memory structure," *Microelectronic Engineering* **41/42**, 511 (1998).
11. C. D. Chen, Y. Nakamura, and J. S. Tsai, "Aluminum single-electron nonvolatile floating gate memory cell," *Appl. Phys. Lett.* **71**, 2038 (1997).
12. A. Nakajima, T. Futatsugi, K. Kosemura, T. Fukano, and N. Yokoyama, "Room temperature operation of Si single-electron memory with self-aligned floating dot gate," *Appl. Phys. Lett.* **70**, 1742 (1997).
13. K. Yano, T. Ishii, T. Sano, *et al.*, "128 Mb early prototype for gigascale single-electron memories," *Proc. IEEE International Solid-State Circuits Conference* (1998), p. 344.
14. W. Chen and H. Ahmed, "Fabrication and physics of ~2 nm islands for single electron devices," *J. Vac. Sci. Technol. B* **13**, 2883 (1995).
15. Y. Ishikawa, N. Shibata, and S. Fukatsu, "Creation of highly-ordered Si nanocrystal dots suspended in SiO₂ by molecular beam epitaxy with low energy oxygen implantation," *Jpn. J. Appl. Phys.* **36**, 4035 (1997).
16. A. Dutta, M. Kimura, Y. Honda, *et al.*, "Fabrication and electrical characteristics of single electron tunneling devices based on Si quantum dots prepared by plasma processing," *Jpn. J. Appl. Phys.* **36**, 4038 (1997).
17. T. Nakanishi, B. Ohtani, and K. Uosaki, "Construction of semiconductor nanoparticle layers on gold by self-assembly technique," *Jpn. J. Appl. Phys.* **36**, 4053 (1997).
18. M. Kawabe, Y. J. Chun, S. Nakajima, and K. Akahane, "Formation of high-density quantum dot arrays by molecular beam epitaxy," *Jpn. J. Appl. Phys.* **36**, 4078 (1997).

19. H. Fujikura, M. Araki, Y. Hanada, M. Kihara, and H. Hasegawa, "Formation of two-dimensional arrays of InP-based InGaAs quantum dots on patterned substrates by selective molecular beam epitaxy," *Jpn. J. Appl. Phys.* **36**, 4092 (1997).
20. C. Lebreton, C. Vieu, A. Pepin *et al.*, "Coulomb blockade effect through a 2D ordered array of Pd islands obtained by colloidal deposition," *Microelectronic Eng.* **41/42**, 507 (1998).
21. M. Mejias, C. Lebreton, C. Vieu *et al.*, "Fabrication of Coulomb blockade devices by combination of high resolution electron beam lithography and deposition of granular films," *Microelectronic Engineering* **41/42**, 563 (1998).
22. C. Wasshuber, H. Kosina, and S. Selberherr, "A comparative study about single-electron memories," to appear in *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.* (1998).
23. K. K. Likharev and A. N. Korotkov, "Analysis of Q_0 -independent single-electron systems," *Proc. Intern. Workshop Computational Electronics* (1995), p. 42.

New Prospects for Terabit Integration

K. K. Likharev

Physics Dept., SUNY at Stony Brook, Stony Brook, NY 11794-3800, U.S.A.

1. Introduction

Semiconductor microelectronics is arguably the most successful technology in the history of our civilization. For several decades, its development has been characterized by exponential ("Moore's Law") progress in several directions, most notably in scaling down silicon transistors and circuits. This remarkable progress serves as a driver for the continuing rapid fall of price per function, enabling the fast progress of digital microelectronics as a whole, and its numerous applications. If this exponential progress were to stop, it would have immediate negative implications for the entire high-tech sector of the present-day global economy, which is why the future prospects of integration are of such importance.

The most authoritative industrial forecast, the National Technology Roadmap for Semiconductors¹ predicts that the exponential increase in integration scale may continue for the next 15 years, leading eventually to the mass production of logic circuits with density up to 180 million transistors per cm^2 and, most importantly, dynamic random-access memories (DRAMs) with a density of almost 20 Gb per cm^2 and integration scale up to 256 Gb per chip.

However, this forecast does not mean that the future is cloudless. First of all, DRAM integration beyond 4 Gb will require fabrication of VLSI circuits with a minimum feature size λ of 100 nm and less, a problem for which there are "no known solutions".¹ Apparently, the further decrease of λ is only possible using either x-ray or e-beam patterning. Both technologies are considerably more complex than the present-day optical lithography, and transfer to them would require a hefty chunk of investment to new fabrication facilities some time in the early 2010s. The semiconductor industry is coming to this critical point considerably weakened by the recent dramatic fall of DRAM prices and hence of the profit margin available for such investment. This weakness is why the transfer to alternative patterning technologies may be delayed until a much later date, and there is a chance it may not happen at all, if the semiconductor technology does not show sufficient ability for innovation. (I believe that the current DRAM production crisis is a result of inadequate innovation, in particular, of sticking to essentially the same main memory idea for too long.)

Second, even if the 100-nm frontier is overcome and fabrication technology at the ~50 nm level (forecast by the Roadmap) proves to be practical, the implementation of 64 Gb and 256 Gb DRAM generations is very much in doubt, mostly due to the problem of scaling down the charge storage capacitors. In fact,

according to the principle of DRAM operation, the storage capacitance C of each memory cell should be sufficient to charge output interconnects all the way down to the sense amplifier, to a voltage high enough in comparison with the amplifier noise — see, e.g., Ref. 2. When the memory cell is scaled down, the capacitance C should remain virtually the same (unless a complex hierarchy of sense amplifiers is used, leading to the loss of effective density). Since 3D tricks with the storage capacitors (trenches, pillars, *etc.*) have already been used, the main current trend is to utilize materials with high dielectric constant ϵ , such as SrTiO_3 ($\epsilon \approx 300$). However, even if the incorporation of these complex materials into the silicon-based technology of DRAM fabrication turns out to be possible, a simple calculation shows that the resulting increase of C will not suffice for the density necessary for 256 Gb DRAM chips, let alone terabit integration.

Recognition of this situation has triggered a wave of research in search of alternative technologies that would enable the current scaling trend to continue after the traditional DRAM approach runs out of steam, in order to reach the terabit integration frontier. During the past few years this research was mostly focused on single-electron devices. These devices (for general reviews see, e.g., Refs. 3, 4) are based on the controllable transfer of single electrons between nanoscale conducting "islands" separated by tunnel barriers. Several digital devices based on this principle have been demonstrated experimentally, and more options have been analyzed theoretically. This research has proven that digital single-electron devices are really possible, can really operate quite reliably and, moreover, may be scaled down to atomic size. The studies have, however, indicated two major problems pertinent to this approach.⁴

The first problem is the requirement of sub-nanometer island size necessary for reliable operation of single-electron devices at room temperature. Though the fabrication of the first single devices of this size by scanning probe techniques was reported recently (see, e.g., Ref. 5), these techniques can hardly be extended to VLSI circuits, because of their very low speed. The second major problem of single-electron logic circuits is their high sensitivity to single charged impurities trapped in the dielectric environment. The electric field of these impurities induces random fractional "background charges" in the islands and as a result causes variations of the device switching thresholds, probably making them unacceptable for VLSI circuits.⁴

A few years ago our Stony Brook group started to look for means to circumvent these problems and open a door to terabit-scale integration. While we did not have much success with *logic* devices, new exciting options have been found in the field of ultradense *memories* and *data storage*. This article contains a brief review of these findings and their possible impact on digital electronics.

2. Crested tunnel barriers

A strong challenge to the currently dominating memory technology, DRAM, might be made by electrically-alterable floating-gate memories (EEPROM, *etc.*⁶).

In addition to being nonvolatile (which is very important, e.g., for mobile systems), floating-gate memories are more scalable than DRAMs, since their intrinsic capacitors need not charge any long external wires, and hence may be shrunk together with the rest of the cell.

The largest apparent drawback of these memories is the slowness of the write/erase operations. Because of this, at present floating-gate structures are used only in read-only memories or "flash" memories for simultaneous writing/erasing large blocks of data,⁶ rather than for the mainstream bit-addressable applications. The reason for this slowness is that in existing floating-gate memories either the write process, or erase process, or both are based on field-enhanced ("Fowler-Nordheim") tunneling through a barrier separating the floating gate from the electron source.⁷ This barrier must have negligible current (corresponding to a gate charge retention time of at least one year) for relatively low voltages applied to the barrier, $V < V_1$. Parameter V_1 characterizes the maximum voltage during data storage, including that created by the stored charge and external voltages applied to write/erase data in other cells of the same row or column ("half-select crosstalk"). On the other hand, in the full-select mode, the applied voltage V_2 should suppress the barrier to such an extent that tunneling current recharges the gate very quickly. In order to keep the memory architecture simple, the ratio V_2/V_1 should be as low as possible (ideally, below 2, making it possible to use one transistor per bit).

The usual uniform SiO_2 tunnel barriers shown in Fig. 1(a) cannot satisfy these two conditions simultaneously. For example, Fig. 2 shows the gate recharging time scale as a function of voltage V for a typical tunnel barrier. The time is determined by the barrier current density, which has been calculated using the standard quasiclassical approximation, in the assumption of the isotropic and parabolic dispersion law for electrons both in the source conduction band and under the barrier, and taking into account the image charge effects.⁷

The results indicate that, for example, a 5-nm-thick barrier of height $U = 3.6$ eV may provide a 3-year retention time ($\sim 10^8$ s) for voltages below $V_1 \approx 3.3$ V, while the write time at $V_2 = 2V_1 \approx 6.6$ V is about 3 ms, far too long for bit-addressable applications. A change in the barrier thickness d to either side only makes the situation worse. A change in the height of the barrier (say, to $U = 3.2$ eV typical for SiO_2) also does not change the situation much.

This relatively weak dependence of barrier transparency on the electric field is due to the fact that the highest part of the barrier, closest to the electron source, is only weakly affected by the applied voltage: $U_{\text{max}}(V) \approx U_{\text{max}}(0)$ — see the dashed line in Fig. 1(a).

Now consider a "crested" barrier with the potential barrier height peaking in the middle and gradually decreasing toward the conducting electrodes, see Fig. 1(b). Figure 2 shows that the current through this barrier changes much faster, so that a voltage change from $V_1 \approx 3.2$ V to $V_2 \approx 5.95$ V $< 2V_1$ decreases the recharging time from 10^8 s to 10^{-8} s. The reason for this dramatic improvement is that in the crested barrier the highest part (in the middle) is pulled down by the electric field very quickly:

$$U_{\max}(V) \approx U_{\max}(0) - eV/2. \quad (1)$$

A similar positive effect on the sensitivity to electric field of thermionic emission (which dominates for barrier heights comparable with $k_B T$) was reported earlier.⁸⁻¹⁰ Moreover, the use of this effect for floating gate memories has been suggested by Capasso *et al.*¹⁰ However, the authors of Ref. 10 considered asymmetric triangular barriers shown in Fig. 1(c). The injection characteristics of such barriers are even better than those of symmetric crested barriers, but only for one current direction (say, "write"). The speed of the reciprocal process ("erase") is low, thus ruling out bit-addressable applications. Of course, this opportunity may be restored by connecting two barriers with opposite barrier slopes in parallel, but this option may be too complex for practical applications.

The implementation of crested barriers is straightforward in composite semiconductors, where the barrier shaping may be achieved with either a gradual change of the layer composition during its epitaxial growth^{8,10} or by modulation doping.⁹ However, the maximum barrier height (conduction band offset) available in these materials is too small to provide sufficient retention time at room temperature. For most prospective wideband materials (SiO_2 , Si_3N_4 , AlN , *etc.*) both these approaches run into fabrication problems; for example for these

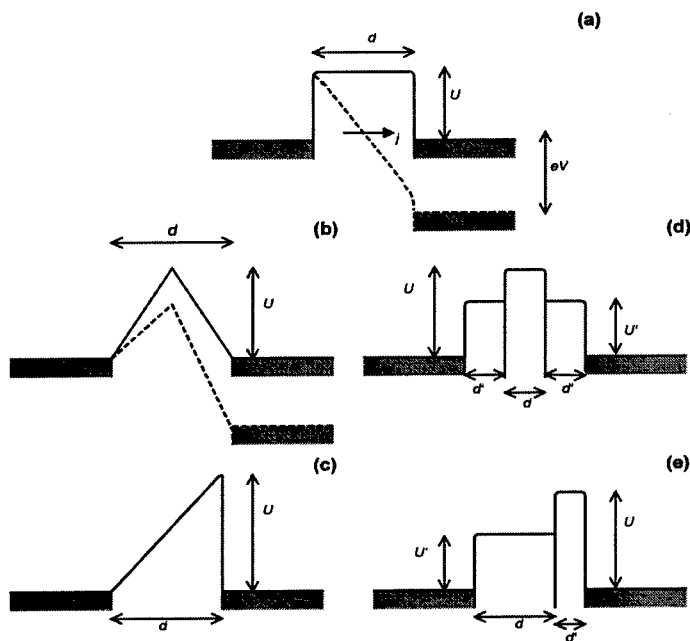


Figure 1. Conduction band edge diagrams of various tunnel barriers: (a) a typical uniform barrier; (b) idealized crested symmetric barrier; (c) idealized asymmetric barrier, (d) crested, symmetric layered barrier, and (e) asymmetric layered barrier. Dashed lines in panels (a) and (b) show the barrier tilting caused by applied voltage V . (After Ref. 12).

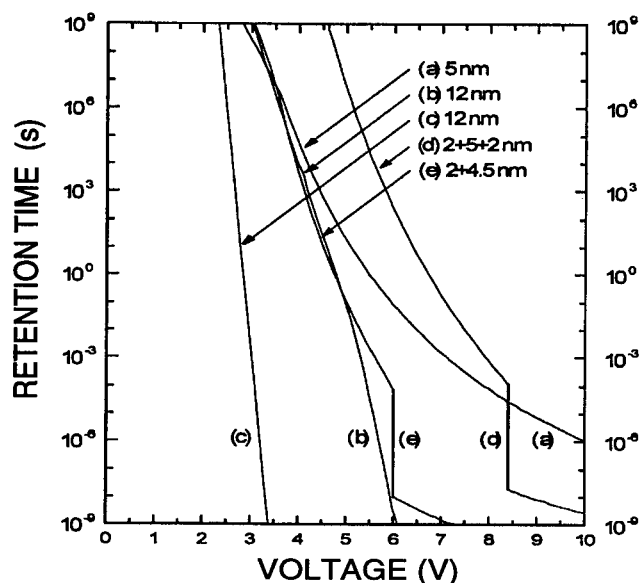


Figure 2. Time scale of floating gate recharging (in seconds) for various barriers, as functions of applied voltage, calculated using the quasiclassical approximation. The curve labeling corresponds to Fig. 1. An isotropic effective carrier mass of $0.2m_0$ was assumed in the electrodes (modeling n^+ -Si). Other parameters are as follows: higher barrier $U = 3.6$ eV, $m = 0.48m_0$, and $\epsilon = 8.5$ (AlN); lower barrier $U' = 2.0$ eV, $m' = 0.2m_0$, and $\epsilon' = 7.5$ (Si_3N_4). (Adapted from Ref. 12.)

materials suitable dopants with shallow levels, necessary for modulation doping, have not yet been found.

Fortunately, there is another possible solution to this problem, which seems much more practical.^{11,12} Both the symmetric and asymmetric barriers shown in Fig. 1(b) and 1(c), respectively, may be reasonably well approximated by "staircase" potential patterns formed in layered barriers (Fig. 1(d) and 1(e)). Calculations of the current density and recharging time have been carried out¹² for the following systems: n^+ -Si/ Si_3N_4 /AlN/ Si_3N_4 / n^+ -Si (trilayer, symmetric barrier, Fig. 1(d)) and n^+ -Si/ Si_3N_4 /AlN/ n^+ -Si (bilayer, asymmetric barrier, Fig. 1(e)), within a broad range of layer thicknesses d and d' . This particular set of materials has been selected since both silicon nitride and aluminum nitride have been successfully deposited on silicon substrates, mostly using a variety of CVD techniques.¹³ Also, for these materials the relevant data, including the conduction band offsets, effective masses, and dielectric constants, have been published.¹³⁻¹⁵

Lines (d) and (e) in Figure 2 show the calculation results for the sets $\{d, d'\}$ providing the lowest ratio V_2/V_1 for the retention time of 3 years and write/erase time of 10 nanoseconds. The sharp current step in each plot is due to the beginning of charge accumulation in a potential dip that is formed at the interface

between the first and second layers as a result of potential tilting by the applied electric field. Beyond the step, direct tunneling through the barrier as a whole is replaced with sequential tunneling via the accumulated free electron layer at the interface. The results show that with the appropriate choice of layer thicknesses, the ratio V_2/V_1 may be below 2 for barriers of both types. Another interesting result is that the advantage of the asymmetric barriers (so evident for the perfectly triangular shape) fades away if the layered implementation is used. This fact, combined with the simplicity of the symmetric barrier option, probably makes it preferable for applications.

One more encouraging result is that the low V_2/V_1 ratio may be combined with a low absolute value of field necessary for fast write/erase: for the symmetric, trilayer barrier it is as low as 6.5 MV/cm. At so low an electric field, the hopping ("Frenkel-Poole") conductance of the nitrides via deep localized states⁷ should not be essential, ensuring high endurance of the barriers under electric stress. Interface trap charging in this system should not be important as well, at least for bit-addressable applications with their low duty cycle. For example, for the cases presented in Figs. 1(d) and 1(e), the traps would discharge through the 2-nm Si_3N_4 layers in less than a nanosecond after the write/erase operation has been completed, making the cell ready for a new cycle.

3. NOVORAM

If confirmed experimentally, the acceleration of Fowler-Nordheim tunneling in layered barriers may have several important applications. First of all, the introduction of crested barriers may turn the usual floating-gate structure of Fig. 3(a) into a unique memory cell that may compete in speed not only with DRAM, but also with power-hungry static random access memories (SRAM). It is also very important that such non-volatile random-access memory (NOVORAM) may be at least as dense as DRAM even at the current technological level.

Moreover, since NOVORAM cells do not need large storage capacitors, the only possible limit to their scaling is set by that of the readout FET. Until very recently, it was believed that silicon FETs could retain useful performance only if their channels were longer than ~ 30 nm. Our recent analysis¹⁶ has shown, however, that *n*-MOSFETs featuring ballistic electron transfer along an undoped (intrinsic) channel, appropriate (dual-gate) geometry, and optimized parameters (source and drain doping $\sim 3 \times 10^{20} \text{ cm}^{-3}$ and gate oxide thickness ~ 2.5 nm) may retain a high degree of control by gate voltage even if their channels are as short as ~ 5 nm. This conclusion finds an implicit confirmation in recent experiments with the first 10-nm-scale transistors fabricated using silicon-on-insulator technology and electron-beam lithography.¹⁷

According to estimates,¹⁶ the readout MOSFET should limit NOVORAM scaling at a density of about $2 \times 10^{11} \text{ bits/cm}^2$. At this point, single-electron devices are a very attractive opportunity, provided that both major problems mentioned in Introduction are avoided.

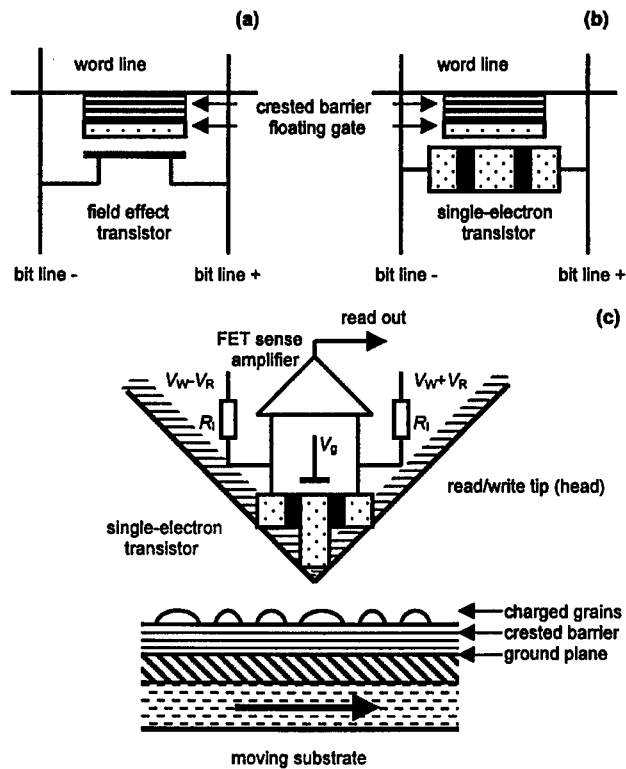


Figure 3. Possible applications of crested barriers: (a) NOBORAM, (b) hybrid SET/FET memory, and (c) system for ultradense electrostatic recording.^{18,21}

4. Hybrid SET/FET memory

Our group has suggested¹⁸ a class of single-electron devices that would circumvent the randomness of background charge. Figure 3(b) shows the schematics of a single cell of the "hybrid SET/FET memory".¹⁸ It is very close to the structure of the NOBORAM cell (Fig. 3(a)), except that the MOSFET is replaced by a single-electron transistor (SET).¹⁹ Since this device is sensitive to random background charges, the readout process should be destructive: it is combined with the "write 1" operation. During this process a small amount of charge $Q = ne$ ($n \approx 30$) is injected into the floating gate. This injection ramps up the electric potential of the floating gate. Due to the periodic transconductance of the SET¹⁹ this ramp up causes a few ($n' < n$) oscillations of its source-drain current. These oscillations are picked up, amplified, and rectified by a FET-based sense amplifier; the resulting signal serves as the output. The main idea behind this design is that the random

background charge will cause only an unpredictable shift of the initial phase of the SET current oscillations, which does not affect the rectified signal. This concept has been verified in recent experiments with a low-temperature prototype of the memory cell.²⁰

Why do we need the readout SET, if a FET is finally used to condition the output signal? Calculations show that the SET preamplifier has very low noise, and thus one FET amplifier/rectifier may serve up to 100 memory cells and hence the associated chip real estate per bit is minor. Another attractive feature of the hybrid SET/FET design is the relatively mild fabrication requirements: if implemented using a silicon-based technology, room temperature operation becomes possible if the minimum feature (transistor island) size is scaled down to ~ 4 nm. This size is much larger than that required for purely single-electron digital circuits. The reason for this considerable relief is that in this hybrid memory the SET is used in an essentially analog mode, not so sensitive to thermally-induced single-electron-tunneling events. Even taking into account the sense amplifiers, decoders, and drivers, the 4 nm design is consistent with the very impressive 10^{11} bit/cm² density. The estimated power density (~ 3 W/cm², mostly in the FET sense amplifiers) also seems quite acceptable.

According to our estimates, the SET/FET hybrid memory may be scaled to a density of $\sim 10^{12}$ bits/cm² (this estimate is for doped-silicon conductors within a SiO₂ matrix, and may be somewhat different for other, more exotic materials). At this point, the floating gate size λ should be about 2 nm, the difference Q of the gate charge in logic states "0" and "1" becomes comparable to the elementary charge e , so that the MOSFET output becomes substantially quantized.¹⁷

At this stage, Nature leaves us apparently with no good choice. On one hand, it would be natural to switch to the direct utilization of the charge quantization (the idea used in most suggestions of digital single-electron devices^{3,4}). However, any attempt to utilize discrete charges to code binary information would be ruined by the randomness of the background charge, mentioned in the Introduction. On the other hand, if we continue to treat the stored charge as a continuum variable, and its discreteness as "noise", the random, Poissonian nature of the integer $n = Q/e$ creates a finite probability of charge transfer by the write/erase process, so that the memory cannot tell "0" from "1". At $\lambda \approx 2$ nm, the probability becomes so large ($\sim 10^{-7}$) that error correction using the system redundancy becomes too costly. Thus, single-electron charging phenomena not only make the SET/FET hybrids possible, but also limit the prospects of their scaling.

Figure 4 summarizes my estimate of the prospects for the development of semiconductor bit-addressable memories. The DRAM development predictions have been borrowed from the SIA forecast.¹ It seems that it will be very difficult to push these memories beyond the 64-Gb integration frontier (density ~ 10 Gb/cm²), mostly due to problems with storage capacitor scaling. This endpoint would leave us with ~ 50 -nm fabrication technologies. Since room-temperature SET/FET hybrids are only feasible starting from ~ 4 nm minimum feature size, this enormous technological gap would be virtually impossible to cross. Fortunately, it appears that NOVORAM may provide a bridge over this gap. Moreover, I expect

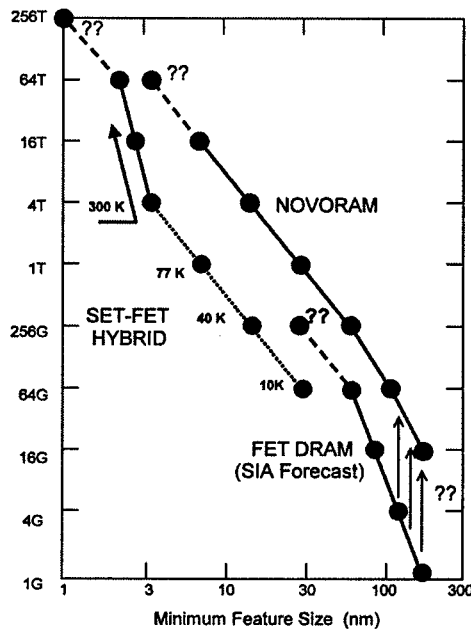


Figure 4. Integration scale expectations for bit-addressable memories. The solid line labeled "DRAM" shows the SIA forecast;¹ the author of this paper is responsible for all other predictions. The bit density is recalculated into the integration scale using the log-linear extrapolation similar to that used in the SIA forecast ($\times 1.4$ chip area increase per generation). Dashed lines indicate problems with no known solutions; dotted line shows SET/FET hybrids which need cooling to the temperature indicated near each point.

that these floating-gate memories with crested barriers may present a serious challenge to DRAMs even at the present technological level, since the non-volatility they offer is very important for low-power electronics applications. If this transition (indicated by vertical arrows in Fig. 4) really happens soon, the gradual technology improvement from one memory generation to the next, which has characterized the last three decades of semiconductor electronics, may continue and eventually bring us terabit-scale integrated circuits.

5. Electrostatic data storage

One more important possible application of crested barriers, when combined with the idea of SET/FET hybrids, is electrostatic data storage — see Fig. 3(c).²¹ The system includes a read/write head with an SET preamplifier loaded on an FET amplifier a few microns apart. (This proximity is necessary to reduce the

interconnect capacitance that is recharged using the SET's high output impedance, which otherwise limits the high read-out bandwidth). The data bits are stored as few-electron charges ($Q/e = n \sim 30$) trapped in nanoscale conducting grains deposited on the top of a crested tunnel barrier. Since the charge is relatively large and is stored in a few (~ 10) grains, their exact shape and location are not important, so the storage medium does not require any nanofabrication: the grains may be deposited, e.g., by simple metal evaporation.

Bit writing is achieved by the application of high voltage V_W in the moment when the head is passing over the specified location (at writing, $V_R = 0$, so that the SET is deactivated and works just as a single conductor delivering voltage V_W to the tip). The voltage suppresses the tunnel barrier and injects the charge into the group of grains. Nondestructive read-out is achieved by the SET activation ($V_W = 0$, $V_R \neq 0$), turning the device into an ultrasensitive electrometer. For this application, the randomness of the background charge is not important, since it may always be compensated by an additional gate voltage V_g , tuned to bias the SET into a point with maximum sensitivity.

Preliminary estimates show that with a 30-nm tip-to-substrate distance (typical for advanced magnetic storage systems), the electrostatic storage system is capable of a density $\sim 10^{11}$ bits per cm^2 (in more traditional units, about 1 Terabit per square inch). This density is two orders of magnitude higher than the best results for magnetic recording of which I am aware, and apparently ~ 30 times higher than the theoretical limit for the magnetic technology. (This limit is imposed by the random thermally-activated jumps of the magnetization of small particles.) In contrast with previous ideas for the implementation of electrostatic data storage (see, e.g., Ref. 24) the use of crested barriers and SET/FET hybrids may provide a very broad bandwidth of both write/erase and read operations, up to ~ 300 Mbit/s per channel, possibly adequate even for this enormous bit density.

The recent experiments at Lucent Technologies²³ may be considered as the first step toward the implementation of this idea. In these experiments a SET electrometer fabricated on a scanning probe was used for the detection of single-electron charges on Si and GaAs substrates. There was no FET amplifier close to the SET output, so that the available measurement bandwidth was very low. However, recently several groups have demonstrated the possibility of broadband SET/FET integration (so far, at low temperatures).^{24,25} It seems that the combination of these achievements with the progress in fabrication of room-temperature SET⁵ and the standard hard disk mechanics technology opens a straightforward way toward the implementation of practical ultradense electrostatic data storage systems.

6. PeT: year 2005 dream system

I believe that the concepts outlined above, especially that of NOVORAM, may flourish in their own right, being incorporated into a broad range of electronic systems. It is interesting, however, to speculate what would be possible if the most

advanced concepts of *logic*, *memory*, and *data storage* were to be merged in a single, stand-alone digital system.

While *density* is the most important figure of merit for memories, the performance of logic circuits is mostly determined by their *speed*. The fastest logic devices that have so far been developed belong to the so-called rapid single-flux-quantum (RSFQ) family.^{26,27} When implemented with mid-submicron design rules, simple digital devices of this family can operate at 750 GHz,²⁸ while VLSI circuits of this family may apparently sustain average clock frequencies above 100 GHz.²⁷ Another substantial advantage of these circuits is their extremely low power consumption, approximately 100 nW per gate operating at 100 GHz. Finally, the technology of fabrication of RSFQ circuits using low-temperature superconductors is rather simple: it is virtually just a subset of the CMOS fabrication process, but with larger minimum feature size.

The main handicap of RSFQ technology is the necessity of deep (helium) refrigeration, which does not allow RSFQ to compete with CMOS for most digital electronic applications. However, in high-performance systems the refrigeration would be only a small fraction of the total cost. As a result, the most powerful driver of RSFQ technology today is the JPL-led HTMT project²⁹ aimed at the eventual development of a *petaflops* computer, i.e. a system capable of sustained performance on the order of 10^{15} floating-point operations per second.

A preliminary design of the RSFQ subsystem of a petaflops computer³⁰ has indicated that its basic unit may look like a 20×20 cm² multichip module carrying 8 superconductor processing elements, a switching network, and a very limited amount (from 4 to 16 Mbytes) of fast superconductor memory. This module could have a peak performance of about 2 Teraflops. Let us see what would it take to use such a module as a core of a stand-alone teraflops-scale computer.

According to the scaling laws of computer design,³¹ in order to sustain a near-peak performance (say, close to 1 Teraflops) on a broad range of practical applications, it should have an associated memory of about 1 Tbyte. This capacity would require as many as 2000 of the most advanced 4 Gb DRAM chips, which should be available by year 2003.¹ However, if by that time the transition from DRAM to NOVORAM has been made (see arrows in Fig. 4), then 64 Gb chips could be available using the same fabrication level ($\lambda = 130$ nm), apparently still within the reach of photolithography. In this case, the number of chips in a teraflops system (with one RSFQ MCM) would be reduced to ~ 128 , which might allow them to be packed on about 16 PC boards fitting into a single rack. Besides that, NOVORAM might allow a faster memory cycle than DRAM, so that memory latency would be easier to hide (using a spectrum of multithreading, pre-fetching, and caching techniques — see, e.g., Refs. 29, 30).

Let us add an electrostatic hard drive system of say 64 disks, 10 in² each. With the recording density of 10^{12} bits/in² (see Sec. 4 above) this system would give an adequate data storage capacity of ~ 80 Tbytes. The resulting computer (including the closed-cycle helium refrigerator for the RSFQ chips) would consume just about 1 kW of power and occupy physical space comparable to one office desk. A crude estimate of the possible cost of such a quasi-desktop teraflops system (when

produced in considerable volume, see below) is about \$100,000, dominated by NOVORAM and hard disk systems rather than the RSFQ processing core.

All these numbers are still in the "personal system" league; this is why I call this concept the "Personal Teraflop", PeT for short. The cited price would be a factor of 5 larger than that of the high-end CMOS workstations expected by year 2005. However, using the SIA predictions¹ the performance of such workstations may be estimated as 20 Gflops at most, limited by CMOS power dissipation. Hence it is reasonable to expect that the PeT approach would provide at least an order of magnitude better price-to-performance ratio.

If such a system becomes commercially available in 2005, what would be the potential market size? Let us estimate one "niche": worldwide, there are at least 1 million professionals in the sciences and engineering (including university faculty, and senior researchers and engineers in corporate and national laboratories). It is reasonable to expect that a sizable fraction will find PeT-level performance useful to their needs, perhaps even indispensable if other workers in the field have access to such a system. If a good fraction of these professionals are compelled to accept an average 5-year delay to accumulate the necessary \$100K budget, we still arrive at a several \$B/yr market "niche", which probably deserves a better name.

Making this naïve estimate, I have ignored another important possibility: the use of a PeT-like system as a group/corporate server. The reader may add other feasible applications. It is my current feeling that this game may be well worth pursuing!

7. Conclusion

To summarize, I believe that floating-gate structures with crested barriers present a remarkable opportunity. These structures may be the basis of unique nonvolatile bit-addressable memories that would be denser and faster than DRAM even at the present technology level, and in addition scalable all the way down to 10^{12} bits/cm². The second important possible application of the crested barriers is electrostatic storage systems with density up to 1 Terabit/in². When implemented, these systems may revolutionize digital memory and data storage technologies. In particular, if combined with novel ultrafast logic circuits, these ultradense systems may bring teraflops computing power right onto your desk in just a few years.

8. Acknowledgments

The author is grateful to his colleagues from Stony Brook and other institutions for numerous useful discussions of the issues in this article. The work on nanoscale silicon-based devices at Stony Brook is supported in part by ONR/DARPA within the framework of Ultra Electronics and Advanced Microelectronics programs. RSFQ work is supported in part by DARPA, NSA, and NASA via JPL within the framework of HTMT project, and also by AFOSR/BMDO and by ONR via HYPRES, Inc.

References

1. *The National Technology Roadmap for Semiconductors, 1997 Edition* (Semiconductor Industry Association, San Jose, CA, 1997).
2. B. Prince, *Semiconductor Memories*, 2nd ed., Chichester, UK: Wiley, 1991; A. K. Sarma, *Semiconductor Memories*, New York: IEEE Press, 1997.
3. K. K. Likharev, "Correlated discrete transfer of single electrons in ultrasmall tunnel junctions," *IBM J. Res. Develop.* **32**, 144 (1988); D. V. Averin and K. K. Likharev, chapter in: H. Grabert and M. H. Devoret, eds., *Single Charge Tunneling*, New York: Plenum, 1992.
4. K. K. Likharev, "Single-electron devices and their applications", to appear in *Proc. IEEE* (1999).
5. J. Shirakashi, K. Matsumoto, N. Miura, and Konagi, "Room temperature Nb-based single-electron transistors," *Appl. Phys. Lett.* **72**, 1893 (1998).
6. W. D. Brown and J. E. Brewer, eds., *Nonvolatile Semiconductor Memory Technology*, New York: IEEE Press, 1998.
7. S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed., New York: Wiley, 1981.
8. C. L. Allyn, A. C. Gossard, and W. Weigmann, "New rectifying semiconductor structure by MBE," *Appl. Phys. Lett.* **36**, 373 (1980).
9. R. J. Malik, T. R. AuCoin, R. L. Ross, *et al.*, "Planar-doped barriers in GaAs by molecular beam epitaxy," *Electron. Lett.* **16**, 837 (1980).
10. F. Capasso, F. Beltram, R. J. Malik, and J. F. Walker, "New floating-gate AlGaAs/GaAs memory devices with graded-gap electron injector and long retention times," *IEEE Electron Dev. Lett.* **9**, 377 (1988).
11. K. K. Likharev, "Novel silicon-based nanoscale devices for terabit memories," in: *GOMAC'98 Digest of Papers* (1998), p. 395.
12. K. K. Likharev, "Layered tunnel barriers for nonvolatile memory devices," *Appl. Phys. Lett.* **73**, 2137 (1998).
13. V. J. Kapoor and K. T. Hankins, eds., *Silicon Nitride and Silicon Dioxide Thin Insulating Films*, Pennington, NJ: The Electrochemical Society, 1987, pp. 7, 23; V. I. Belyi, L. L. Vasilyeva, A. S. Ginovker, *et al.*, *Silicon Nitride in Electronics*, Amsterdam: Elsevier, 1988, pp. 148, 162; S. Strite and H. Morkoç, "GaN, AlN, and InN: a review," *J. Vac. Sci. Technol. B* **10**, 1237 (1992).
14. J. T. Wallmark and J. H. Scott, "Switching and storage characteristics of MIS memory transistors," *RCA Review* **30**, 335 (1969).
15. V. W. L. Chin, T. L. Tancley, and T. Osotchan, "Electron mobilities in gallium, indium, and aluminum nitrides," *J. Appl. Phys.* **75**, 7365 (1994); V. M. Bermudez, T. M. Jung, K. Doverspike, and A. E. Wickenden, "The growth and properties of Al and AlN films on GaN (0001)-(1×1)," *J. Appl. Phys.* **79**, 110 (1996).
16. F. Pikus and K. Likharev, "Nanoscale field-effect transistors: an ultimate size analysis," *Appl. Phys. Lett.* **71**, 3661 (1997).

17. L. Guo, E. Leobandung, and S. Y. Chou, "A room-temperature silicon single-electron metal-oxide-semiconductor memory with nanoscale floating-gate and ultranarrow channel," *Appl. Phys. Lett.* **70**, 850 (1997).
18. K. K. Likharev and A. N. Korotkov, "Ultradense hybrid SET/FET dynamic RAM: feasibility of background-charge-independent room-temperature single-electron digital circuits," *Proc. 1995 ISDRS*, Charlottesville, VA, 1995, p. 355.
19. D. V. Averin and K. K. Likharev, "Coulomb blockade of tunneling, and Coherent oscillations in small tunnel junctions," *J. Low Temp. Phys.* **62**, 345 (1986);
K. K. Likharev, "Single-electron transistors: electrostatic analogs of the dc SQUIDS," *IEEE Trans. Magn.* **23**, 1142 (1987);
T. A. Fulton and G. D. Dolan, "Observation of single-electron charging effects in small tunnel junctions," *Phys. Rev. Lett.* **59**, 109 (1987).
20. C. D. Chen, Y. Nakamura, and J. S. Tsai, "Aluminum single-electron nonvolatile floating gate memory cell," *Appl. Phys. Lett.* **71**, 2038 (1997).
21. K. K. Likharev and A. N. Korotkov, "Analysis of Q_0 -independent single-electron systems," in: *Abstr. Int. Workshop Comput. Electron*, Tempe, AZ, 1995, p. 42;
A. N. Korotkov and K. K. Likharev, " Q_0 -independent single-electron systems," *VLSI Design* **3**, 201 (1997).
22. R. C. Barrett and C. F. Quate, "Charge storage in a nitride-oxide-silicon medium by scanning capacitance microscopy," *J. Appl. Phys.* **70**, 2725 (1991).
23. M. J. Yoo, T. A. Fulton, H. F. Hess, *et al.*, "Scanning single-electron transistor microscopy: imaging individual charges," *Science* **276**, 579 (1997).
24. E. H. Visscher, J. Lindeman, S. M. Verbrugh, P. Hadley, and J. E. Mooij, "Broadband single-electron tunneling transistor," *Appl. Phys. Lett.* **68**, 2014 (1996).
25. J. Pettersson, P. Wahlgren, P. Delsing, *et al.*, "Extending the high-frequency limit of a single-electron transistor by on-chip impedance transformation," *Phys. Rev. B* **53**, R13272 (1996);
B. Starmark, P. Delsing, D. B. Haviland, and T. Claeson, "Noise measurements of single electron transistors using a transimpedance amplifier," *Ext. Abstr. ISEC'97*, Berlin, 1997, p. 391.
26. K. K. Likharev and V. K. Semenov, "RSFQ logic/memory family," *IEEE Trans. Appl. Supercond.* **1**, 3 (1991).
27. K. K. Likharev, "Superconductors speed up computation," *Phys. World* **10**, 33 (May 1997).
28. W. Chen, A. V. Rylyakov, V. Patel, J. E. Lukens, and K. K. Likharev, "Superconductor digital frequency divider operating at 750 GHz," *Appl. Phys. Lett.* **73**, 2817 (1998).
29. G. Gao, K. Likharev, P. Messina, and T. Sterling, "Hybrid technology multithread architecture," *Proc. 6th Symp. Frontiers Massively Parallel Comput.*, Los Altos, CA: IEEE Comp. Soc. Press, 1996, p. 98;

- T. Sterling, "A hybrid technology multithreaded architecture for petaflops computing," CACR, Caltech, Pasadena, CA, 1997, available at <http://htmt.cacr.caltech.edu/Overview.html>.
30. M. Dorojevets, P. Bunyk, D. Zinoviev, and K. Likharev, "RSFQ computing: the quest for petaflops," in this book.
31. D. Patterson and J. Hennessy, *Computer Architecture. A Quantitative Approach*, 2nd ed., San Francisco: Morgan Kaufmann, 1996.

Data Storage — Is the End of the Bit Near?

A. V. Nurmikko

Division of Engineering, Brown University, Providence, RI 02912 USA

H. Goronkin

Motorola Phoenix Corporate Research Laboratories, Tempe, AZ 84284 USA

1. Introduction

Closely tied to the technological explosion that is vernacularly known as the "information age", advances in high density storage mirror the furious pace that is the trademark of today's microelectronics-driven computer technology. Whether of permanent, archival forms of magnetic/optical/electronic storage, or volatile and fast random access microelectronic chips, the storage capacities have been increasing typically by at least one order of magnitude every half a dozen years. Such exceptional rates of growth imply a correspondingly diminishing size of the basic element in a memory, i.e. the bit cell. This scaling leads to a question of the possible physical limits in storage density in the foreseeable future, and calls for alternative strategies in terms of innovative new approaches, including the basic physical and material limits that define the action in a bit cell. Examining some of the parameters for such fundamental obstacles and their implications defines the theme of this article.

There are many current flavors of high density storage that take advantage of electronic, magnetic, and optical properties of solids. Technologically dominant among these are the magnetic hard drives, random-access memories, and optical disks, all of which have reached a high degree of maturity and sophistication. Each of these technologies is quite distinct and shares few, if any, common features with another, either from the standpoint of the defining basic physical processes or the system architecture. In the following we focus specifically on issues that relate to the magnetic hard drives and optical disks, to evaluate critically the inherent limitations that are now appearing on the horizon for each, and attempt to envision solutions to these limitations. There are other areas, especially from storage materials point of view, which will not be touched upon here. These areas include the interesting contemporary work on ferroelectrics, for example. In addition to focusing on basic microscopic magnetic and optical processes in material heterostructures that define the storage media, we also prophesize that "hybrid" approaches will be particularly important in the future in cross-fertilizing storage concepts from one class of memory systems to the next. As one contemporary example, we consider ongoing efforts to marry magnetic storage with random access memory (MRAM).

The structure of this article is as follows. In Section 2 we outline the principal challenges facing high density magnetic recording media, from the viewpoint of basic micromagnetic considerations. Section 3 features some issues of optical disk technology, emphasizing magneto-optical storage and the recent demonstrations of MRAMs. Finally, Section 4 provides a summary canvas for spontaneous speculation of offbeat ideas for futuristic storage concepts.

2. Limits for magnetic recording media – the end of the bit?

- *Nature as an eraser*

Any commercial advertisement for personal computers worth its salt today will boast of the product performance by quoting the speed of the processor and storage capacity of the (nonvolatile) memory, i.e. the hard drive (HD). While the basic Winchester-type architecture, inherited from the 1970's, is still recognizable in a contemporary 2.5 inch, 8 Gbit unit, both the recording medium and the head have undergone a remarkable evolution. The head and the thin-film ferromagnetic medium are, of course, coupled in terms of the overall performance of the system while imposing constraints on each other. We will, however, dispose of the head in this section by noting that the next generation of heads is still likely to feature an inductive means of writing a bit, but with the benefits of a (giant) magneto-resistive sensor as a highly versatile and sensitive submicron pick-up element.

Much of the approximately 60% annual growth in the storage capacity of HDs can be attributed to the development of relatively high-coercivity small-grain thin film Co-based ferromagnetic (FM) recording media. The 1 Gb/in² storage density milestone has already been reached in commercial products, media designers are next faced with the challenges posed by the 10 Gb/in² goal during the next couple of years, while the quest for the "holy grail" of 100 Gb/in² looms much cloudier on the horizon.

Figure 1 shows a high resolution TEM micrograph of the grain structure of a typical CoCrTa sputtered thin film on a Cr underlayer, highlighting the segregation of nonmagnetic Cr within the grain boundaries of the Co composition map for exchange isolation.¹ Figure 2 shows a magnetic force microscope image of such recording media in action, with submicron recorded tracks.² Typically, the actual cell bit size must be significantly (perhaps a thousand-fold) larger than the individual grain size of the polycrystalline thin (10–30 nm thick) film. This bit size reduces the stochastic media noise (due to the random micro-crystalline orientation and magnetic coupling) to a practical level. From a micromagnetic point of view, it is important that the grain boundaries be non-magnetic to prevent magnetic (exchange) coupling between grains. While the ~10–30 nm grains now being studied in the laboratory do interact through the long range dipole interaction, they can be viewed as small semi-independent magnetic particles provided that the underlying magnetocrystalline anisotropy is large enough.

A fundamental limit in the stability of a small (single domain) magnetic particle against thermal fluctuations is known as the "superparamagnetic" limit.³

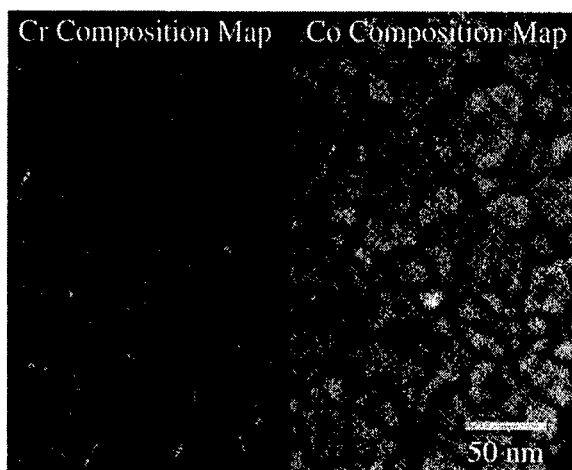


Figure 1. High-resolution energy-filtered TEM images of $\text{CoCr}_{12}\text{Ta}_4/\text{Cr}$ media showing the Cr composition (left panel) and Co composition (right panel) maps.¹

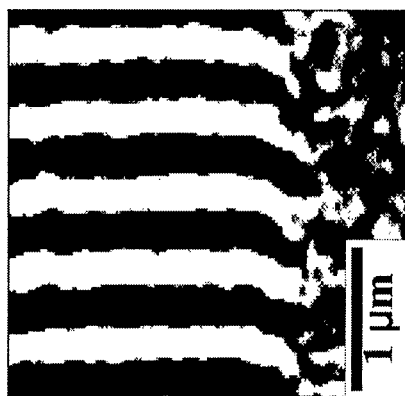


Figure 2. Written bit cell pattern in CoCrTa media, imaged by magnetic force microscopy.²

A practical measure of this limit is given by the ratio of the magnetic anisotropy energy of the crystal (or magnetic particle in case of shape anisotropy being dominant) to thermal energy, $K_u V/kT$ where K_u is an energy density and V is the particle volume. According to the simplest micromagnetic models, the lifetime of a magnetized particle against thermally-induced spontaneous magnetization "flipping" scales as $\tau \sim \exp[K_u V/kT]$. For pure hexagonal cobalt $\tau \sim 1$ year for an ideal spherical particle with diameter $D = 10$ nm, but only $\tau \sim 1$ second for $D = 6$ nm, illustrating the profound impact of the approaching superparamagnetic limit. Hence the only apparent alternative for homogeneous polycrystalline thin film

media is to look for materials with large values of K_u , together with some further "tailoring" of the microscale grain texture. While transition metal/rare earth alloys of high coercivity such as CoSm offer some hope, it is generally agreed⁴ that this type of uniform thin film recording media will be unable to support a storage density of 100 Gbit/in². For this bit density, the individual grain size would have to be very small, <10 nm, making it susceptible to the thermal fluctuations that lead to magnetic instability, hence destroying the memory content of the bit cell.

A perhaps obvious but challenging escape hatch from the superparamagnetic conundrum is the design and synthesis of patterned media, composed of an array of small, magnetically truly independent but highly uniform magnetic particles. Thus, for example, a 50 nm periodicity for a two dimensional array of "dots" of the same diameter would in principle yield a density of > 200 Gbits/in², with each magnetic (e.g. Co) particle still stable against thermal fluctuations. Clearly, lithographic techniques are available today to create such structures, as shown in Fig. 3.⁵ The detailed micromagnetic studies of submicron particles, in general, are today an active area of basic research. The appealingly simple concept represented in Fig. 3 involves, however, challenges at two very different levels. First, minute imperfections and shape anisotropies can "pin" the magnetization within a given dot, giving it both static and dynamical magnetic fingerprints that are specific to the particular bit cell, both in terms of the signal and its noise. Secondly, a patterned medium will require new write/read head architectures and concepts compatible with the dot geography, including the design of the head servo and tracking systems for a safe, reliable and fast access to the disk. Furthermore, we want to emphasize that lithographic patterning is only one avenue for creating magnetic nanostructure arrays. Deposition of a thin magnetic film on an artificially or naturally patterned substrate template is a concept actively pursued in the "nanomagnetism" community. A variety of "natural templates" are being

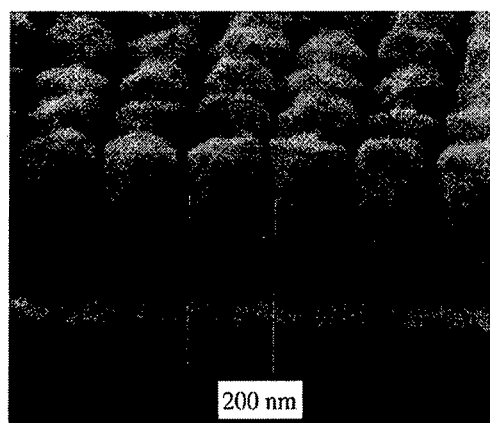


Figure 3. A high density array of submicron nickel magnetic particles fabricated by optical holographic lithography.⁵

investigated, ranging from nanoporous media (e.g. alumina)⁶ to biological organisms (such as bacteria).⁷ Another future direction in the synthesis of two-dimensional arrays of nanomagnetic particles might rely on self-assembly of specific materials (e.g. quantum dots in strained layer semiconductor heterostructures) that are doped with magnetic elements or upon which a magnetic film is deposited with preferential growth on specific topographic features (e.g. quantum dot facets). Such self-assembled techniques could also eventually lead to fully 3D architectures of magnetic ultra-high density volume memory, to which access would be provided by electrical or optical means.

- *Breaking the speed limit*

Aside from the question of engineering high-speed random access to a future nanostructured high-density thin film medium, the fundamental limits to speed in the actual writing of a bit, i.e. actuating magnetization reversal, loom increasingly large on the horizon. Present projections are that switching rates within the hard drive medium of approximately 1 GHz and beyond will be required by industry already within a few years at the level of a single bit. Today, in a typical case of a multidomain bit, the switching time in a practical magnetic field delivered by the head (~1000 Oe) is limited to about ≥ 1 ns by the much studied domain wall motion. Recent work by time-resolved magneto-optical spatial imaging techniques⁸ has shown how this type of magnetization reversal proceeds by a complex pathway, with domain wall nucleation and propagation proceeding from the edges inward of a 10 μm scale thin film particle. In the case of a small single-domain magnetic particle, however, it is possible to induce in principle the magnetization reversal in the so-called coherent rotation regime, provided that the transient magnetic field pulse is short and intense enough, and that the "magnetic viscosity" (damping) is low.⁹ Even with the help of high speed (> GHz) microstriplines to carry the inducing current pulse, it is difficult to generate sub-nanosecond transient magnetic field pulses that are intense enough to study and exploit the limits of magnetization switching in standard high coercivity ferromagnetic storage media. Classically, coherent magnetization dynamics are described by the Gilbert-Landau-Lifshitz governing equations¹⁰ and bear a close analogy to the equations of motion for electron spin precession and nuclear magnetic resonance. Recently, some off-the-beaten-path attempts have been made in the laboratory to test the switching speed limit in thin ferromagnetic films. Siegmann *et al.* used the high energy beam from a linear accelerator to create a pulsed high magnetic field (~4 kOe) of ~10 ps in duration within a thin Co film.¹¹ Sample damage prevented a detailed analysis of the transient magnetization vector $\mathbf{M}(t)$ behavior. However, it appears that some form of coherent rotation of the magnetization vector was accomplished on a comparable ultrafast timescale. Elsewhere, Ju *et al.* have experimented with a rather different concept, exploiting fast "photomagnetization" effects induced by ultrashort laser pulses.¹² In this instance the magnetic system was a NiFe/NiO ferro-magnetic/antiferromagnetic bilayer, where the interfacial exchange coupling of the FM and AF layers creates an internal unidirectional exchange bias field that shifts the hysteresis loop of the

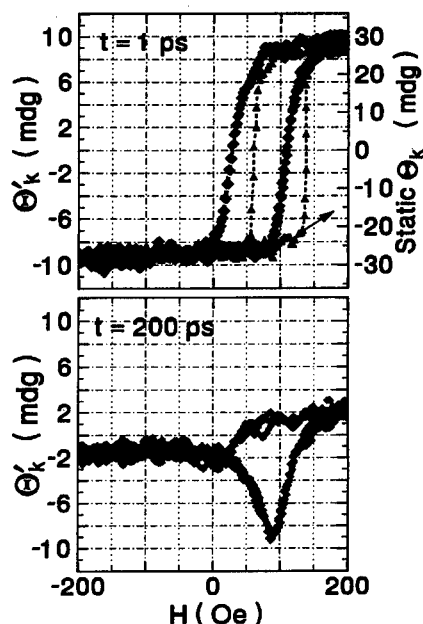


Figure 4. Optical control of magnetization in an exchange biased FeNi/NiO (10 nm/40 nm) bilayer. The figure compares the instantaneous hysteresis loop, recorded at the arrival of an ultrashort laser pulse, with another "snapshot" taken after 200 ps, when large changes in the system magnetization are complete.¹²

FM layer.¹³ In the transient experiment, the absorbed photons from a picosecond laser pulse created a "hot" spin population near/at the interface, effectively "shorting out" the built-in exchange field (for a duration of approximately 100 ps, the spin-lattice relation time). This field modulation activated a "torque" on the magnetization vector \mathbf{M} in the permalloy layer. Figure 4 shows the impact of such "switching" through the dramatically modified appearance of the photonicallly switched hysteresis loops at two selected time intervals relative to the optical excitation pulse, as well as the static magnetization characteristics (vertical axis corresponds to the measured magneto-optical Kerr rotation, which is proportional to magnetization).

3. When in doubt, think hybrid

The present storage technologies are sharply focused in their implementation in that specific system functions are assigned to particular physical actuators, which rely on the optimization of a given dominant physical phenomenon. Hence the main "degrees of freedom", that is, magnetism, electronics, and photonics are viewed and exploited as distinct and well separated entities. At some future point

it seems inevitable that both the storage medium as well as the access to it (write/read tools) will have to be configured so as to couple and integrate these primary components at a basic phenomenological level in some monolithic way. Much innovation will be required (and created) by such hybrid approaches, from which next generations of storage technologies are likely to emerge. Here we consider two contemporary schemes, both under active research and development, which serve as examples in this direction.

- *High density magneto-optical storage*

Magneto-optical (MO) storage is part of a family of optical disk technologies where information is deposited on a thin film by local thermal modification of its appropriate physical property (such as magnetization or crystallinity). Coherent laser light also promises the ultimate high-density storage in its own right through holography. Since the material requirements for fully 3D holographic storage have so far proven to be exceptionally difficult, we focus here on 2D thin films (such as the opaque MO and phase change media). In connection with MO recording, in particular, research and development is presently under way to look for an optical bridge to "conventional" HD technology.

One means for bridging MO and HD storage combines near-field optical techniques with a "Winchester drive"-like access to reading/writing of an optical disk by light below the conventional diffraction limit. Figure 5 illustrates schematically two versions of the near field approach, where the effective light source or optical detector lies within a fraction of the wavelength from the optical disk surface. Either the phenomenon of total internal reflection is exploited in a precisely crafted, high numerical aperture lens scheme (similar to the solid immersion lens¹⁴ employed in highest resolution optical microscopes), or an actual

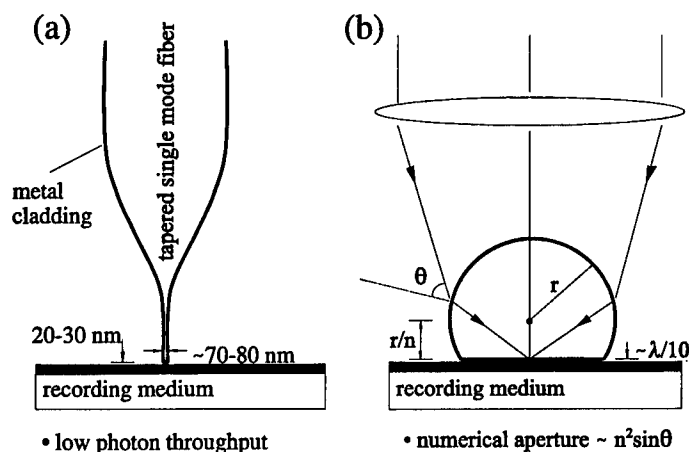


Figure 5. Schematic of approaches to optical sub-diffraction microscopy and recording: (a) the tapered fiber; (b) the solid immersion lens.

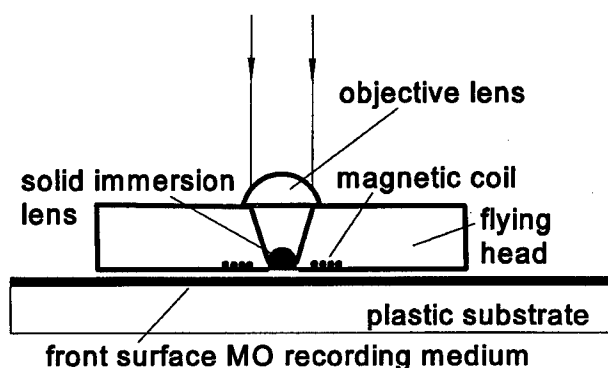


Figure 6. Schematic of an optical "Winchester" flying head, utilizing near field, evanescent light waves for magneto-optical recording/readout.¹⁶

physical aperture is used to define a sub-100 nm optically active spot within the evanescent field of the aperture for illumination and collection. At the moment, the latter approach has been realized with tapered optical fibers, shown in Fig. 5(a), that have too low a photon throughput for storage applications. For visible light, the near field regime requires the placement of the optical source/detector within a few hundred Å from the disk surface – a distance comparable to the height of the present flying head in a magnetic hard drive. With a compact semiconductor laser as the source, this scheme raises the prospect of incorporating the laser and accessory optics within the head of a Winchester drive so as to replace the magnetic head write coil or sensor (or both) by a sub-diffraction optical stylus. A U.S.-based company (Terastor) recently announced a commercial implementation of one near-field scheme, with laser light delivered by fiber and micro-optical components to the flying head, as shown in Fig. 6. The product, scheduled for release in late 1998, is a 20 Gb industry standard 5.25" form factor storage product.¹⁶ We also note that in terms of the ultimate performance of near field optical schemes, an elegant demonstration has been made recently, in which optical interference effects further enhance the spatial resolution and sensitivity of an "apertureless" arrangement, for a potential storage capacity up to several hundred Gbits/in².¹⁷

A parallel evolution that will further increase such near field storage capability is the upcoming availability of blue and violet semiconductor lasers (in the 400 nm wavelength range and below) that are expected to emerge from research laboratories to the commercial marketplace within a year or so.¹⁸ Apart from the obvious and already assured application of these sources to next generation CD-ROMs, DVDs, and related optical disk products, the above mentioned optical Winchester drive scheme and other sub-diffraction approaches to ultra-high density optical storage will directly benefit from the new GaN lasers. Simple wavelength scaling alone would suggest that an individual bit size in the near field could be on the order of 800 Å in the solid immersion lens scheme,

suggesting a disk capacity approaching the Tbit/in² range! However, formidable obstacles present themselves in the design of a fast access scheme at such extreme storage density. One can envision parallel processing schemes with modern optoelectronics with holographic elements and spatial light modulators that actuate and control multiple laser beam access to the disk.

The near optical field concept is also conducive to possible realization of a large enhancement in the light-matter interaction between the source and the target, provided that there is sufficient local feedback between the electromagnetic (light) and the electronic oscillator (e.g. surface plasma resonance of the medium). For example, the planar microcavity that will be formed by the highly reflecting optical surfaces between the head and the medium could be exploited to greatly enhance the efficiency of energy delivery from a write beam, while enhancing the sensitivity of the readout process. Perhaps the ultimate realization of this coupled source/receiver/media idea is the possibility of integrating the MO medium as part of the feedback (mirror) structure of a planar, short vertical cavity surface emitting laser. Finally, we refer in passing to the question of the ultimate speed in MO recording, already touched upon in Section 2. With the anticipated availability of short pulse (sub-picosecond) blue and violet diode lasers sometimes in the next decade, it may become feasible to write a magnetic bit of information *nonthermally*. Nonthermal writing is here defined as a transient, nonequilibrium event, where the excitation of the spin and electronic degrees of freedom, inducing an irreversible change in the state of magnetization in the target, takes place *prior* to the energy relaxation to the lattice via electron-phonon and spin-lattice interaction (i.e. heating). Since the effective heat capacity of the spin bath is much smaller than that of the lattice, the energy budget for inducing a magnetic transition in the system purely electronically (by light) should be at least an order of magnitude smaller than in the present "heat-gun" approaches to optical recording.

- *Magnetic random access memory*

A novel approach that brings "magnetism-to-a-chip" to memory applications is embodied by the magnetic random access memory (MRAM). Research and development of MRAMs, in which the marriage of magnetism and electronics has been under study for at least a decade, has now reached an active phase at leading industrial laboratories. Progress in the field has been fueled in large part by the discoveries of new magnetotransport phenomena, notably the giant magnetoresistance (GMR) in layered ferromagnetic as well as other heterogeneous metals. The essence of the GMR effect is summarized in Fig. 7, both in the "spin-valve"¹⁹ and the "magnetotunnel junction" (MTJ)²⁰ configurations. In the case of the spin-valve, the ferromagnetic layers are exchange coupled through a thin (say 10–20 Å) nonmagnetic metal layer. Lateral transport through the sandwich shows resistance changes as large as 10% between the parallel and anti-parallel orientations of the magnetization vectors within the two FM layers (composed typically of Ni_{0.80}Fe_{0.20} permalloy, or NiFeCo ternary). Similarly, in the magnetotunnel junction, even larger resistance changes can be measured due to

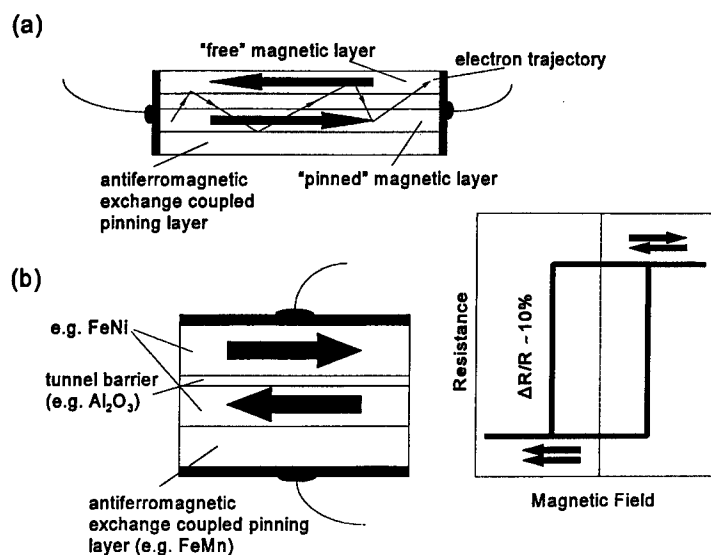


Figure 7. Basic principles for implementing giant magnetoresistance via (a) the spin valve, and (b) the magnetotunnel junction approaches. In (a) the resistance of the multilayer is highest when the magnetization of the topmost "free magnetic layer" opposes that of the bottom "pinned" magnetic layer. In (b) the tunneling current is minimized when the ferromagnetic electrodes have their magnetizations in opposite directions.

differences in the (spin dependent) tunneling current for the opposing magnetization orientations of the FM electrodes. In a bit cell device, the magnetization of one of the FM layers is usually fixed, or pinned, by the exchange interaction with an immediately adjacent antiferromagnetic underlayer, this coupled system being magnetically biased to possess unidirectional anisotropy during the synthesis of the multilayer material. Because the exchange interaction is of very short (atomic) range, such a multilayer spin valve or MTJ heterostructure is only a few hundred Å thick. In particular, its substrate can be based on Si or related materials, allowing the integration of the basic magnetic memory cell with standard microelectronics circuitry, including CMOS technology. In this manner, the marriage of magnetics and electronics is offering fresh opportunities for the development of a matrix addressable, nonvolatile on-the-chip memory, as a complement to the volatile SRAMs and DRAMs.

Examples of these types of memory cell device structures and illustrations of their performance are shown in Fig. 8. Magnetic information for each cell is written by Faraday induction and probed via the GMR effect in each (submicron) cell by orthogonal "word" and "sense" (strip) lines, whose current amplitudes are chosen for maximum economy and impact in the writing process, while

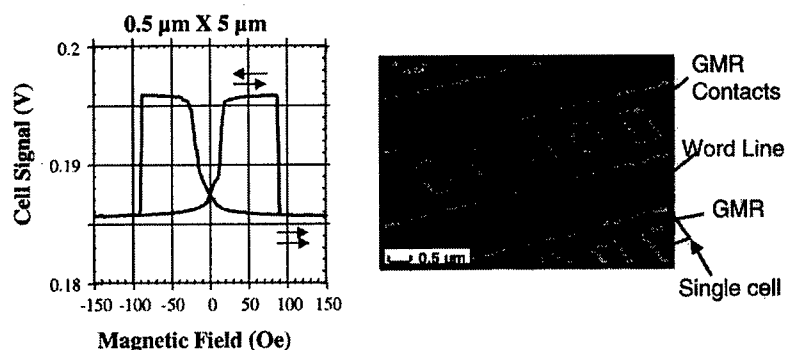


Figure 8. Illustration of the architecture and performance of a spin-valve MRAM cell.²¹

optimizing the sensitivity/noise for the read operation, respectively.²¹ Prototype spin valve MRAMs with capacity of 256 Kb have been recently demonstrated at Honeywell, with access times comparable to the volatile RAM technology. Future issues with these types of MRAMs include the scaling of the cell size down to and below the $0.25 \mu\text{m}$ regime, given some of the challenges of magnetization behavior and stability (Section 2). In any event, the incipient MRAM technology is likely to find a strong technological niche for many on-board applications in low-power electronic units such as cellular phones. From a broader perspective, its development serves as a useful model for the "hybrid" strategy of mixing magnetism with modern microelectronics, with the inclusion of photons possibly yet to come (e.g. in the spirit of the nonequilibrium ultrafast magnetization control referred to above). We also note active work underway to embed ferromagnetic constituents within compound semiconductors,²² as well as ferromagnetic metal-semiconductor multilayers²³ that promise novel magnetic sensors and storage concepts based on spin-polarized electronic transport.

4. Wild blue yonder

As evidenced by other articles in this volume, many offbeat yet fundamental scientific concepts can be envisioned as forming an important, perhaps essential, component in the strategic portfolio for future microelectronics, including high density storage. Several of these concepts may appear currently as being both unrealistic and futuristic in their technological implementation anytime soon; yet important first pieces of groundwork are being laid through sophisticated basic research. Single electron devices, molecular electronics, and areas of "spin electronics" may indeed one day form the basis of wholly new concept families for storage and computing, for example, where the coherent (wavefunction) states of matter perform the data processing/storage functions (quantum computing and

related concepts). From the standpoint of this article, spin coherent states are particularly relevant and of interest for such schemes, although the hurdles put up by nature are formidable. In a typical ferromagnetic metal, for example, the high itinerant electron (and spin) density leads to ultrashort spin coherence times ($<10^{-12}$ s), rendering such materials most likely unsuitable for information capture in quantum coherent states. In general, electron spin relaxation in condensed matter, especially crystalline semiconductor or insulating media, tends to be fairly rapid at room temperature (< 1 ns), due to the spin-orbit interaction among the Bloch states. In the limit of high purity silicon, the isolated spins on particular dopant (shallow donor) isotopes may have very long lifetimes ($>> 1$ sec) at cryogenic temperatures. As an alternative to using such single, very weakly interacting spins as a storage or computational node, one can also envision collective spin states, e.g. in nanostructured semiconductors or their hybrids with organic molecules. An array of exchange coupled self-assembled quantum dots with few electrons per dot may offer the prospect of designing a macroscopic, quantum mechanically correlated coherent state, i.e. the ultimate ferromagnetic medium.

References

1. J. E. Wittig, T. P. Nolan, C. A. Ross, *et al.*, "Chromium segregation in CoCrTa/Cr thin films for longitudinal recording media," *IEEE Trans. Magn.* **34**, 1564 (1998).
2. E. N. Abarra, G. N. Phillips, I. Okamoto, and T. Suzuki, "DC erasure and demagnetizing fields on written bits in high density longitudinal media," *IEEE Trans. Magn.* **34**, 1621 (1998).
3. See, for example, A. Aharoni, *Introduction to the Theory of Ferromagnetism*, Oxford, UK: Clarendon Press, 1996.
4. M. H. Kryder, W. Messner, and L. R. Carley, "Approaches to 10 Gbit/in² recording," *J. Appl. Phys.* **79**, 4485 (1996).
5. M. Farhoud, M. Hwang, H. I. Smith, *et al.*, "Fabrication of large area nanostructured magnets by interference lithography," *IEEE Trans. Magn.* **34**, 1087 (1998).
6. K. Liu and C. L. Chien, "Magnetic and magnetotransport properties of novel nanostructured networks," *IEEE Trans. Magn.* **34**, 1021 (1998).
7. C. J. Smith, M. Field, C. J. Oakley, *et al.*, "Organizing nanometer scale magnets with bacterial threads," *IEEE Trans. Magn.* **34**, 988 (1998).
8. M. R. Freeman, W. K. Hiebert, and A. Stankiewicz, "Time-resolved scanning Kerr microscopy of ferromagnetic structures," *J. Appl. Phys.* **83**, 6217 (1998).
9. L. He and W. D. Doyle, "A theoretical description of magnetic switching experiments in picosecond field pulses," *J. Appl. Phys.* **79**, 6490 (1996).
10. R. Kikuchi, "On the minimum of magnetization reversal time," *J. Appl. Phys.* **27**, 1352 (1956).
11. H. C. Siegmann, E. I. Garwin, C. Y. Prescott, *et al.*, "Magnetism with picosecond field pulses," *J. Magn. Magn. Mater.* **151**, L8 (1995).

12. G. Ju, A. V. Nurmikko, R. F. C. Farrow, *et al.*, "Ultrafast optical modulation of an exchange biased ferromagnetic/antiferromagnetic bilayer," *Phys. Rev. B* **58**, 19842 (1998).
13. See, for example, A. P. Malozemoff, "Mechanisms of exchange anisotropy," *J. Appl. Phys.* **63**, 3874 (1988).
14. B. D. Terris, H. J. Mamin, D. Rugar, *et al.*, "Near field optical data storage using a solid immersion lens," *Appl. Phys. Lett.* **65**, 388 (1994).
15. See M. A. Paesler and P. J. Moyer, *Near-Field Optics*, New York: Wiley, 1996.
16. Terastor news release, June 1998; see www.terastor.com.
17. Y. Martin, S. Rishton, and H. K. Wickmarasinghe, "Optical data storage at 256 Gbits/in²," *Appl. Phys. Lett.* **71**, 1 (1997).
18. S. Nakamura, M. Senoh, and S. Nagahama, "InGaN/GaN/AlGaIn-based laser diodes with modulation-doped strained-layers," *Appl. Phys. Lett.* **72**, 211 (1998).
19. B. Dieny, V. S. Speriosu, S. S. P. Parkin, *et al.*, "Giant magnetoresistance in soft ferromagnetic multilayers," *Phys. Rev. B* **43**, 1297 (1991).
20. See, for example, W. Gallagher, S. S. P. Parkin, Y. Lu, *et al.*, "Microstructured magnetic tunneling junctions," *J. Appl. Phys.* **81**, 3741 (1997).
21. E. Y. Chen, S. Tehrani, T. Zhu, *et al.*, "Submicron spin valve magnetoresistive random access memory cell," *J. Appl. Phys.* **81**, 3992 (1997).
22. N. Akiba, F. Matsukura, A. Shen, *et al.*, "Interlayer exchange in (Ga,Mn)As/(Al,Ga)As/(Ga,Mn)As semiconducting ferromagnetic trilayer structures," *Appl. Phys. Lett.* **73**, 2122 (1998).
23. D. J. Momsma, R. Vlutters, T. Shimatsu, *et al.*, "Development of the spin valve transistor," *IEEE Trans. Magn.* **33**, 3495 (1997);
Y. B. Xu, E. T. Kernohan, M. Tselepi, *et al.*, "Single crystal Fe films grown on (100) InAs by molecular beam epitaxy," *Appl. Phys. Lett.* **73**, 399 (1998).

Vertically Integrated SRAM

Marco Mastrapasqua

Bell Labs, Lucent Technologies, Murray Hill, NJ 07974, U.S.A.

Gerhard Hobler

Institute of Solid State Electronics, Univ. of Technology, A-1040 Vienna, Austria

Enrico Sangiorgi

DIEGM, University of Udine, Udine, I33100 Italy

1. Introduction

Static random access memory (SRAM) cells are commonly used as embedded memory because they are fast, dissipate low power, and are easy to use since they do not require the overhead circuitry of DRAM or flash EPROM.¹ Furthermore, SRAM can be made with smaller additional process cost than DRAM or FLASH. When volatility is not an essential requirement, SRAM is in direct competition with DRAM. In order to overcome SRAM's main disadvantage there is a strong effort to decrease its cell size: only one transistor and one capacitor are required for a DRAM cell, whereas the SRAM needs six transistors of which two are PMOS. One viable approach is vertical integration. One early solution was to use polysilicon resistors as the load and fabricate them on top of the NMOS transistors. Subsequently, the need to decrease the power and increase the stability to noise and soft errors forced the use of a full CMOS cell design. In this case, a successful design formed the two PMOS load transistors in the polysilicon layer above the four NMOS transistors in the substrate.²

In this article we suggest a method to further increase the vertical integration of an SRAM cell without sacrificing the self-aligned nature of the MOS process and without "substantially" increasing the number of process steps.

2. Vertical Integration of an SRAM

A schematic circuit representation of a six-transistor SRAM is sketched in Fig. 1. Notice that in the case of the *n*-type drive transistors M1 and M2, the drain of one transistor is connected to the gate of the other and *vice versa*. Normally such a connection is made with a metal line contacting the drain and the gate polysilicon.¹ A much more compact arrangement, which eliminates completely the need of an external drain-to-gate connection, can be achieved by building the drive transistors on opposite sides of the gate oxide, as shown in Fig. 2.³ In this vertical

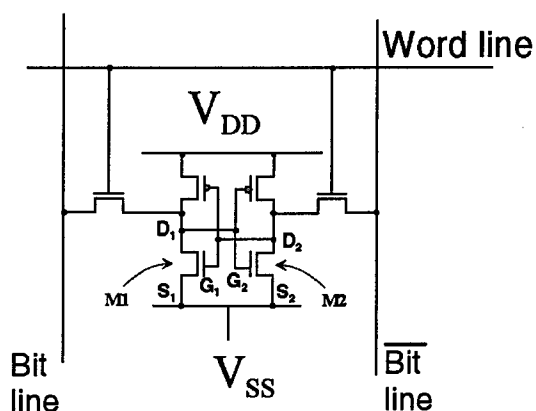


Figure 1. Schematic representation of a CMOS SRAM cell.

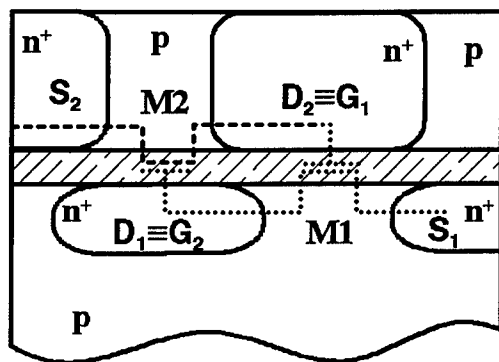


Figure 2. Schematic representation of the vertical integration of the two NMOS of the flip-flop of Fig. 1. The shaded area is a gate oxide separating the two channels. The drain of the M1 transistor, D_1 , acts as the gate of M_2 ; while the drain of the M2 transistor, D_2 , acts as the gate of M_1 .

integration scheme, the connection between the drain of one transistor and the gate of the other is not necessary because the drain replaces and acts as the gate of the other transistor.

To complete the SRAM cell, access transistors and load will be connected to the drain of M1 and M2. The load can be a polysilicon resistor or, as in Fig. 1, a PMOS transistor, either bulk or thin-film. For the scope of this paper we will concentrate on the vertical integration of the two NMOS drive transistors forming the flip-flop; in principle, however, a similar vertical integration scheme could also be applied to the two PMOS devices acting as loads. Eliminating the drain-to-gate connections combined with the vertical integration structure will definitely reduce

the cell size. An exact quantification of the advantage in reducing the cell size is premature at this stage because it will depend on the complete layout, type of isolation, and contacts used.

Two major problems facing the practical realization of the structure sketched in Fig. 2 are the epitaxy of the multilayer structure and the desired doping profiles. As for the former problem, recent progress indicates that selective epitaxial lateral overgrowth (ELO) might provide the necessary buried oxide with the very narrow thickness required by a gate insulator.⁴ In ELO, single crystal material is grown from oxide-masked seed windows. After reaching the top of the opening, the growth proceeds laterally as well as vertically, forming silicon on the insulator. However, to date ELO has been demonstrated for masking oxide thickness only down to 80 nm.⁵ Therefore the application of this technique to grow single crystal material over much thinner oxides has not been proven yet. Moreover, it has not been demonstrated that the quality of the oxide-laterally grown silicon interface is sufficient for MOSFET operations.

In this article we concentrate on the second problem, how to introduce the doping of the structure. In the next section we present the proposed process together with process simulations results. Then, in Section 4, we present the simulated device and SRAM cell characteristics. Finally, we draw some conclusions in Section 5.

3. Process description

The proposed process for obtaining the desired vertical structure of Fig. 2 is independent of the method used to obtain the starting multi-layer material. In the following we assume that a 6 nm thick silicon dioxide layer separates the two silicon layers, and that the epitaxial Si (epi-Si) layer is 250 nm thick. The bulk and epi-Si layers are *in-situ* boron doped to concentrations of 10^{18} and $1.7 \times 10^{18} \text{ cm}^{-3}$, respectively. The upper epi-Si layer has to be doped slightly higher than the bulk-Si layer in order to compensate *n*-type dopants that are unavoidably introduced during the doping of the n^+ regions of the lower MOSFET.

The n^+ regions of the MOSFETs are doped in a self-aligned manner using one and the same mask structure. This scheme is advantageous not only because it will reduce the processing cost, but more importantly because it offers the possibility to self-align the top transistor to the bottom one, which is essential for the correct operation of the cell. We assume both the mask length and opening to be 250 nm. The n^+ layers in both the epi-Si and the bulk-Si should have high concentrations at the Si/SiO₂ interface, while keeping the concentrations of *n*-type dopant at the opposite side of the oxide negligible or sufficiently low to be compensated by the boron background doping. One possible way to achieve such a strong concentration gradient above and below the gate oxide is by channeling implantations along $\langle 110 \rangle$ directions.

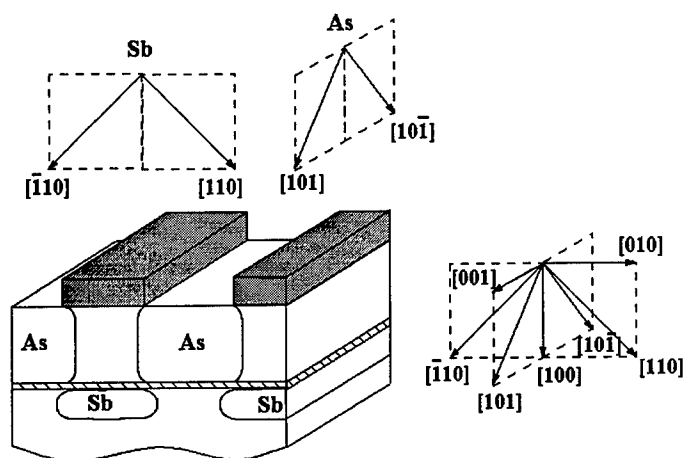


Figure 3. Proposed self-aligned implants to simultaneously dope the upper and lower layer. All implants are along $\langle 110 \rangle$ channeling directions. Antimony is implanted in the $[\bar{1}10]$ and $[110]$ directions at 60 keV, while arsenic is implanted along the $[10\bar{1}]$ and $[101]$ directions at 6 keV. It is essential that the thickness of the upper epitaxial silicon layer and the mask length be the same.

The n^+ regions in the epi-layer are implanted with 10^{15} cm^{-2} 6 keV As ions along the $[101]$ and $[10\bar{1}]$ directions — see Fig. 3. These directions have no component in the $[010]$ direction, which results in a small lateral extension of the 2D dopant distributions, while the small stopping power of the $\langle 110 \rangle$ channels allows many of the dopants to penetrate close to the oxide (see Fig. 4(a)). The slope of the As profile near the oxide is much steeper than in conventional higher-energy tilted implantation. This slope allows a lower thermal budget to diffuse the dopants to the Si/SiO₂ interface and therefore reduces lateral diffusion. The n^+ regions in the bulk-Si are produced by $5 \times 10^{13} \text{ cm}^{-2}$ 60 keV Sb implantations along the $[110]$ and $[\bar{1}10]$ directions (Fig 3). At this energy a large fraction of the channeled ions penetrates through the epi-layer and is dechanneled at the gate oxide. Dechanneling causes a much larger number of ions stopped per unit path length and results in a correspondingly larger concentration of Sb ions below the oxide. As can be seen from Fig. 4(b), a ratio of 25 can be achieved with the current parameters. The use of Sb as a dopant for the bottom layer is occasioned by its smaller diffusivity and higher ratio of channeled/random implantation range. The lateral components of the $[110]$ and $[\bar{1}10]$ directions make it possible to use the same mask structure as for the As implant to produce the laterally displaced n^+ doping of the bulk-Si, provided that the epi-Si layer thickness equals the mask length. The mask has to stop all of the implanted ions and yet avoid excessive shadowing that would corrupt the bulk-Si n^+ regions — requirements met by a 95 nm thick mask for our process parameters. All implantations are performed at high temperature to avoid target amorphization that would inhibit the channeling effect. Finally, the dopants are activated by a 5 minute furnace anneal at 1000 °C.

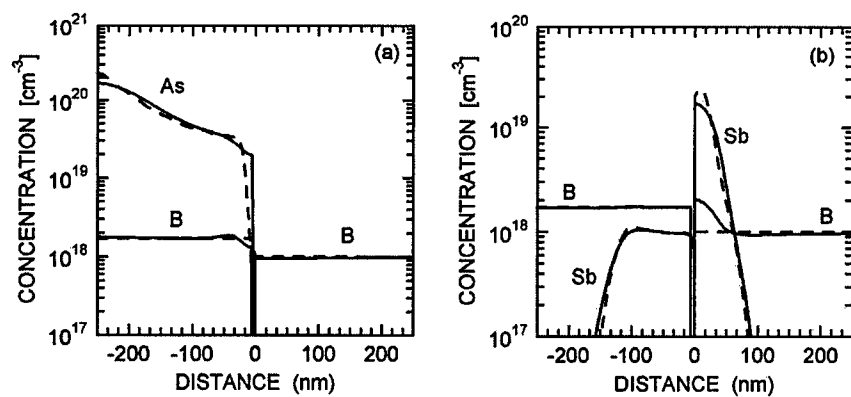


Figure 4. Doping profile as-implanted (dashed lines) and after anneal at 1000°C for 5 minutes (solid lines): a) doping along the section AA of Fig. 5; b) doping along the section BB of Fig. 5.

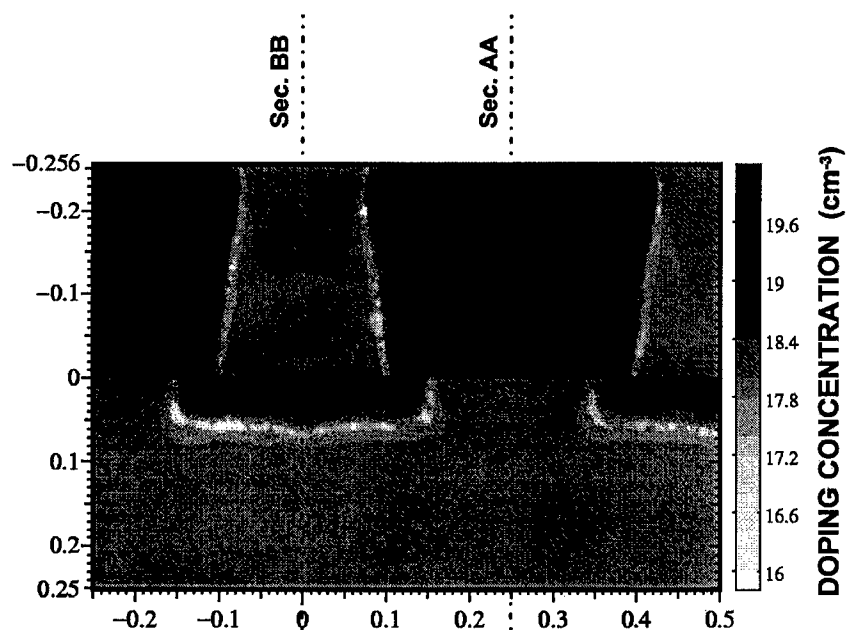


Figure 5. Two-dimensional doping profile obtained by simulating the implants along the $\langle 110 \rangle$ directions with the Monte Carlo simulator IMSIL and the furnace anneal at 1000°C for 5 minutes with the partial differential equation solver PROPHET. The grayscale corresponds to the logarithm of the absolute value of the net doping concentration.

Figure 4 shows the as-implanted profiles (dashed lines) as simulated with the implant simulator IMSIL.⁶ The diffused profiles as simulated with PROPHET⁷ are shown with solid lines. As can be seen, the 1000 °C anneal for 5 minutes is sufficient to bring a high concentration of As to the oxide. At the same time there is only moderate diffusion of the Sb profile. Figure 5 shows the post-anneal 2D dopant distribution, which is used for the device simulation described below. Obviously, since the proposed process uses implants along channeling directions, it is necessary that the top epi-Si layer crystallize in a known orientation. Both these conditions could be satisfied by obtaining the multi-layer structure by ELO. However, as we mentioned, other techniques able to provide a similar layered structure could be valid solutions.

4. Device simulations

Using the 2D drift-diffusion simulator PADRE⁷ and starting from the 2D doping profile of Fig. 5, we have simulated the static and dynamic characteristics of the vertical SRAM assuming, for simplicity, a passive load. The nominal transistor channel length $L = 0.25 \mu\text{m}$ (as set by the mask parameters), while the transistor width $W = 1 \mu\text{m}$. The chosen value of the load was 100 k Ω .

The inset of Fig. 6 shows the simulated transfer characteristics of the bulk-Si (M1) and epi-Si (M2) NMOS transistors. The results indicate that M1 and M2 exhibit very similar behavior, with negligible threshold voltage difference. The drive current capability of M2 is slightly lower, due to the higher doping level (lower channel mobility) of the M2 channel region.

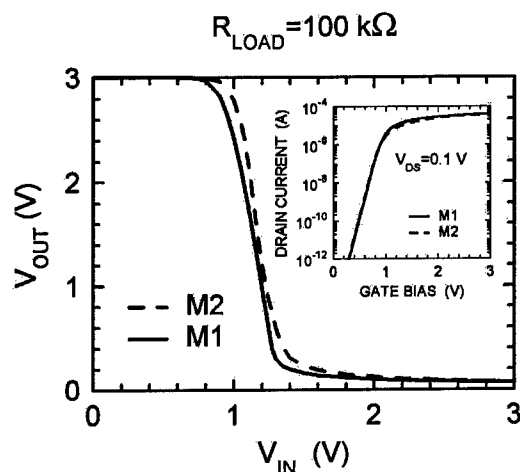


Figure 6. Input-output characteristic of the vertical SRAM cell of Fig. 5 simulated with a drift-diffusion solver using 100 k Ω as a load. Inset shows simulation of the drain current vs. gate bias for both transistors.

Next, each drain was connected with the lumped load element and the corresponding gate was stepped in voltage to obtain the I/O characteristics of the two inverters forming the SRAM cell. The results (Fig.6) show that, thanks to the symmetry of M1 and M2, the inverter characteristics are also very symmetric. In fact, the higher logic threshold shown by the M2 inverter is not due to any asymmetry of the two MOSFETs. Instead, when the gate of M2 ($G2 \equiv D1$ in Fig.2) is stepped up, M1 is on, due to the fact that its gate ($G1$) is also the drain of M2, which is still at high voltage. Thus current flows along $D1 \equiv G2$ and a considerable voltage drop is established due to the relatively low doping level of $D1 \equiv G2$. As a consequence, a higher gate voltage is required by the M2 inverter to switch. To minimize such an asymmetrical behavior, the process should maximize the doping levels of the Sb-doped n -type regions.

Finally, the dynamic characteristics of the cell were simulated assuming only the intrinsic device capacitance, embedded in the drift-diffusion simulations, and ignoring any parasitic contributions. Thus the results are indicative only of the intrinsic behavior of the vertical structure. The simulations were performed by alternatively applying a positive bias pulse to the M1 (M2) gate with the cell in the M1 = off, M2 = on (M1 = on, M2 = off) state, and then checking for the correct switching of the cell. The results, shown in Fig. 7, indicate that an input voltage pulse of only 130 ps is sufficient to switch the memory cell in both configurations. Notice that the switch is slower when the pulse is applied to the M2 transistor. This difference is again due to the relatively low doping level of the M2 gate region, whose series resistance effectively slows down the switching of the cell, and indicates the need for doping engineering optimization to produce highly symmetrical drive transistors.

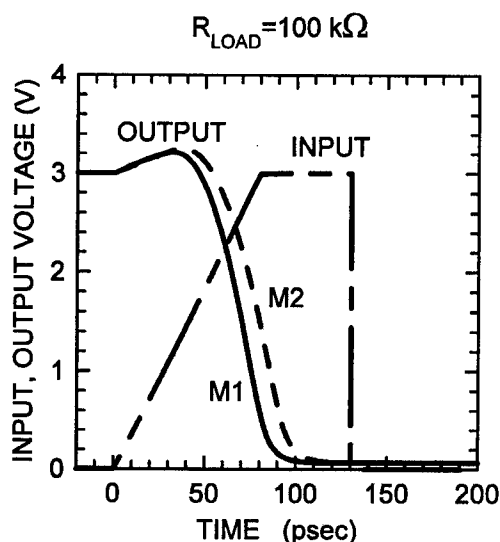


Figure 7. Transient switching characteristics of the vertical-SRAM cell of Fig. 5 simulated with a drift-diffusion solver using 100 k Ω loads.

5. Conclusions

In this article we have proposed a novel vertical SRAM structure. By building the two MOS drive transistors on the opposite sides of a thin gate oxide, the drain of one acts as the gate of the other and *vice versa*. This vertical integration eliminates the need of a metal line to connect the drain to the gate and thus has the potential of increasing the packing density. To realize the proposed vertical SRAM structure we suggest using selective epitaxial silicon lateral overgrowth followed by implants along $\langle 110 \rangle$ channeling directions. The use of implants along channeling directions allows doping the source and drain region for both top and bottom transistors with the same mask. Monte Carlo simulations of the channel implants combined with process and device simulations confirm that the structure behaves like an SRAM.

Although practical implementation of this device has not yet been demonstrated, we believe that such a vertical integration scheme may become a viable technique to increase device density without unduly increasing process cost.

6. Acknowledgments

The authors wish to acknowledge helpful discussion with M. Alam, P. Diodato, A. Ghetti, C. King and the support of M. Banu, S. Hillenius, K. Ng, and C. Rafferty.

References

1. Betty Prince, *Semiconductor Memories*, 2nd ed., New York: Wiley, 1995.
2. M. Ando, T. Okazawa, H. Furuta, *et al.*, "A 0.1 μ A standby current bouncing-noise-immune 1 Mb SRAM," *Tech. Digest Symp. VLSI Circuits* (1988), p. 49.
3. M. R. Pinto, private communication.
4. J. P. Denton and G. W. Neudeck, "Fully depleted dual gated thin film SOI PMOSFETs fabricated in SOI islands with an isolated buried polysilicon backgate," *IEEE Electron Dev. Lett.* **17**, 509 (1996).
5. J. A. Friederich, M. Kastelic, G. M. Neudeck, and C. G. Takoudis, "The dependence of silicon selective epitaxial growth rates on masking oxide thickness," *J. Appl. Phys.* **65**, 1713 (1989).
6. G. Hobler, "Monte Carlo simulation of two-dimensional implanted dopant distributions at mask edges," *Nucl. Instrum. Meth. B* **96**, 155 (1995).
7. M. R. Pinto, D. M. Boulton, C. S. Rafferty, *et al.*, "Three-dimensional characterization of bipolar transistors in a submicron BiCMOS technology using integrated process and device simulation," *IEDM Tech. Digest* (1992), p. 923.
8. M. R. Pinto, "Simulation of ULSI device effects," in: *Proc. 3rd Intern. Symp. ULSI Sci. Technol.*, Pennington, NJ: Electrochemical Society, 1991, pp. 43-51.

5 Electromagnetic Systems: From Microwaves to the Visible

This chapter covers photonics and analog microwave electronics. This rapidly developing area has seen a number of recent advances on many fronts. New coherent sources are being developed, extending the wavelength range into the visible blue/violet regions on the one side, and mid/far infrared on the other. Moreover, phased antenna arrays are gaining in importance as the coherent sources of millimeter and sub-millimeter waves. An overview of several thrusts in these directions is provided by Yoon-Soo Park and Max Yoder, whose paper on electromagnetic systems opens the chapter.

A thorough review of the status and trends in III-nitride materials and devices is provided by Manijeh Razeghi, who sees the 21st century as the "final frontier" for these material systems, which had seen so much development in the last decade. These are, of course, the necessary material ingredients for systems of immense commercial importance. Materials issues associated with widegap semiconductor devices are also discussed by several other authors in this chapter. At the other end of the spectrum, there is an article by Qing Hu and coworkers discussing the efforts to extend the operation of intersubband unipolar semiconductor lasers in to terahertz region. Still longer wavelengths belong to the realm of semiconductor analog electronics. A paper describing a new approach to RF modeling without a pre-conceived equivalent circuit model is presented by Serge Luryi.

Optical filters, routing and switching elements are the subject of intense current research. This chapter presents several thrusts in this area, ranging from the discussion of polymer optical interconnects by Louay Eldada to the discussion by Jeff Young of two-dimensional photonic lattices. One of the pioneers of photonic bandgap engineering, Eli Yablonovitch, presented an inspiring talk at Embiez on the novel concept of optical code-division multiplexing. As described by C.F. Lam and Yablonovitch in this chapter, an optical CDMA communication system can deliver a throughput of up to one terabit per second.

Electromagnetic Systems Advances

Yoon Soo Park and Max N. Yoder

Electronics Division, Office of Naval Research, 800 N. Quincy St., Arlington, VA
22217-5660

1. Introduction

Enabling technological advances underway in several areas show great promise of impacting electromagnetic (E/M) systems design. Among those advances are crystal growth of low defect and large area wide bandgap semiconductors, new device structures and extremely high performance integrated circuits, both analog and digital. These advances chiefly impact multifunctional E/M systems.

Conventional phase-shift-steered antenna arrays are severely limited in their ability to transmit or receive signals exhibiting large time-bandwidth products and in their ability to simultaneously radiate or receive multiple simultaneous beams — especially when those beams are capable of independently expressing frequency, beamshape, elevation, bearing, power, and signal modulation characteristics. To simultaneously achieve this degree of versatility requires true-time-delay beamsteering technology. Over the past several years, various optical delay line approaches have been investigated to accomplish this, but these approaches have been bulky and extremely costly. Moreover, it is difficult to use them to achieve 0.5 ps differential time delay between adjacent elements and such time resolution is necessary to steer an 18 GHz signal with 1 degree beam pointing accuracy.

Alternative approaches to optically derived time delay are accomplished by micro-electromechanical system (MEMS) modules, but while these are ideally suited for obtaining the small differential time delays, the total delay of several nanoseconds needed from one end of the antenna array to the other end of the array effectively defeats the purpose of MEMS. This paper will illustrate *inter alia* how the combination of MEMS incremental delay devices with direct digital synthesis and digitally generated delay is most efficacious in accomplishing the desired beam steering capability in the smallest and lowest cost package.

2. Lateral epitaxial overgrowth

Lateral epitaxial overgrowth (LEO) has been known for some time.¹ Only recently, however, has its efficacy been proven. The first viable demonstration of LEO was in the synthesis of gallium nitride (GaN) on both silicon carbide (SiC) and sapphire substrates.² Figure 1 illustrates the approach for an intended growth of GaN on silicon. Here it is seen that two different levels of mask must be

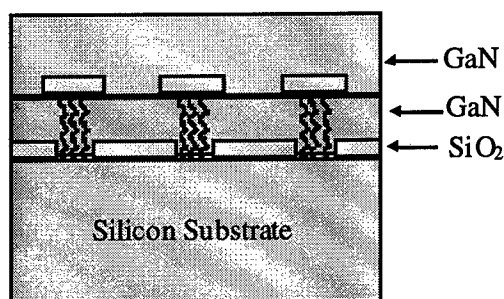


Figure 1. Lateral Epitaxial Overgrowth (LEO).

provided if all defects are to be eliminated. Gallium nitride has been grown by LEO over an entire 2" diameter sapphire substrate. Since GaN tends to rotate when nucleating on sapphire, and this rotation can be different in different nucleating windows, the coalescence of two lateral GaN films may produce a grain boundary. The LEO of GaN has also been demonstrated over an entire SiC 1.25" 6H substrate. In this case there is true epitaxy (no rotation, but nearly 4% lattice mismatch) and coalescence lines are absent. The efficacy of the LEO technique is that the strain in the 1 micrometer wide window is not extended; as the GaN grows out laterally over the silicon dioxide mask, the GaN neither nucleates on the mask nor is constrained by it. At the 1080 °C growth temperature, the mask is largely plastic. Thus the GaN grows strain-free. In a continuous (non-LEO) epitaxial growth of GaN on SiC, there is one misfit dislocation every 25 lattice sites. In LEO there is one misfit dislocation every 10 μm (at the coalescence line). Thus the overall strain is much much less than 1% and can easily be accommodated without a coalescence line. The dislocation density is thought to be near zero as defects can be found neither with transmission electron microscopy nor hot etch solutions that would otherwise delineate them.

If GaN can be synthesized by LEO on a silicon substrate, then the potential exists that virtually all semiconductors for which there are no native substrates can be grown nearly dislocation free over large silicon wafers. When overgrowths are of sufficient thickness, the original silicon substrate and mask may be etched away. Mercury cadmium telluride (MCT) is an example of a material system and its IR detector applications that could potentially benefit from LEO. If MCT defect density could be reduced 1000-fold, then IR detector response speed could be much improved, mechanical choppers eliminated in the long wavelength region, and viable performance extended to 15 micrometers. The 6.2-Angstrom material system is another example where LEO may provide better materials. Finally, ultra high speed logic materials such as gallium indium arsenide (GaInAs) may be grown economically and nearly defect-free on large area silicon wafers. The availability of GaInAs in large areas could have a large impact on ultra high speed logic circuits such as those required for high-performance analog to digital (A/D) converters and direct digital synthesizers (DDS).

3. Low parasitic heterojunction bipolar transistors

About 25 years ago the silicon industry realized that while bipolar transistors were generally faster than CMOS devices, they did not scale as effectively for making integrated circuits. The reason behind this lack of scaling for the bipolar devices was the intrinsic inherent parasitic base-collector capacitance directly below the base Ohmic contacts. Recently, this problem has been overcome.³ Using a surrogate substrate and backside processing virtually eliminates all of the parasitic base-collector capacitance of a heterojunction bipolar transistor (HBT). Low parasitic HBTs have exhibited f_{\max} greater than 500 GHz. It is expected that this technology will provide the first 1-THz transistors and from these 100 GHz logic is feasible. Approaches other than flip-chip transferred substrate are also available to reduce the base-collector parasitic capacitance. While the initial versions of 100-GHz logic are anticipated to be in III-V semiconductor (LEO) material, it is possible that silicon-germanium technology may approach this performance level. If so, costs can be expected to be even more reasonable.

Lateral epitaxial overgrowth (LEO) technology described in this article is expected to drive down the cost of the required III-V semiconductor materials.

4. High performance integrated circuits

The low parasitic HBT has already demonstrated rather remarkable performance in simple integrated circuits. One such circuit is a low-pass amplifier that performs from dc to 50 GHz. Figure 2 illustrates this amplifier. These same transistors with different circuitry have recently demonstrated record breaking performance from dc to 350 GHz. Finally, a high-performance digital circuit in the form of a master-slave D flip-flop has operated at a 47 GHz clock speed. It is illustrated in Fig. 3. With a change of load resistors, this circuit is expected to perform at clock speeds exceeding 80 GHz. It is from circuitry similar to this that direct digital synthesizers (DDS) will be developed.

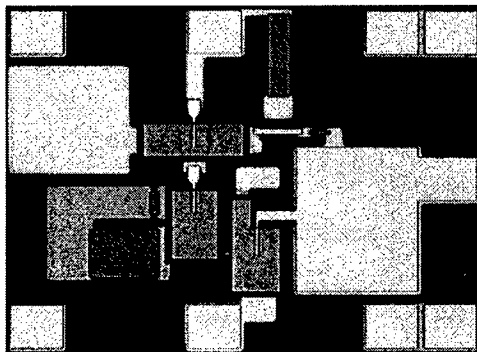


Figure 2. Darlington/cascode feedback amplifier dc to >50 GHz, 10 dB.

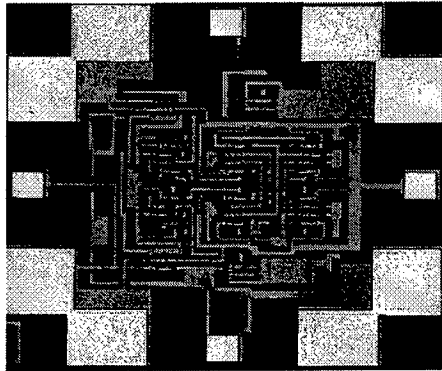


Figure 3. Transferred-substrate HBT ICs in development: master-slave D flip-flop with an 85 GHz clock (design target).

5. Direct digital synthesizers and beam steering

Figure 4 depicts the overall concept for using a direct digital synthesizer for three integrated functions: sinusoidal synthesis, phase and frequency modulation, and course beam steering. A 100-GHz master stable oscillator drives the system. It is distributed first to a "vernier device" and then to the DDS. Phase and frequency modulation codes, as well as beam steering data, are fed to the DDS. The output of the DDS goes to a broadband, linear, efficient semiconductor-based wide-bandgap power amplifier and then to the individual antenna elements. Not shown are the driver amplifiers with integral gain (amplitude) control functions. Amplitude modulation may best be applied in this fashion rather than digitally.

The vernier device is an element wherein the time delay can be controlled over the range of 0 to 10 ps. A phase shifter could be used as it would be a single frequency phase shifter, but this would probably be too expensive. The ideal solution for the vernier is a MEMS device wherein various delay lines can be switched to obtain this delay in 0.1 ps increments. An unlimited number of coarse 10 ps delay increments is obtained by each cycle of the 100 GHz clock with a count down circuit in the synthesizer. This arrangement will provide beam steering accuracy of 0.2 degrees and maximum offboresight steering of 90 degrees (endfire). If the antenna is required to steer a maximum of 45 degrees off boresight, then this corresponds roughly to 9.5 bit accuracy in a conventional phase-shift-steered antenna. Since the MEMS delay devices are very small, they can be implemented in a tandem manner such that one unit controls the beam steering while the other unit can compensate for spatial variations in antenna element placement or feedline path length. To appreciate the magnitude of this problem, in an 18 GHz array, the misplacement of one antenna element by as little as 9 micrometers is equivalent to feeding that element with a signal whose beam direction is 1 degree different than that of adjacent elements. Mutual coupling

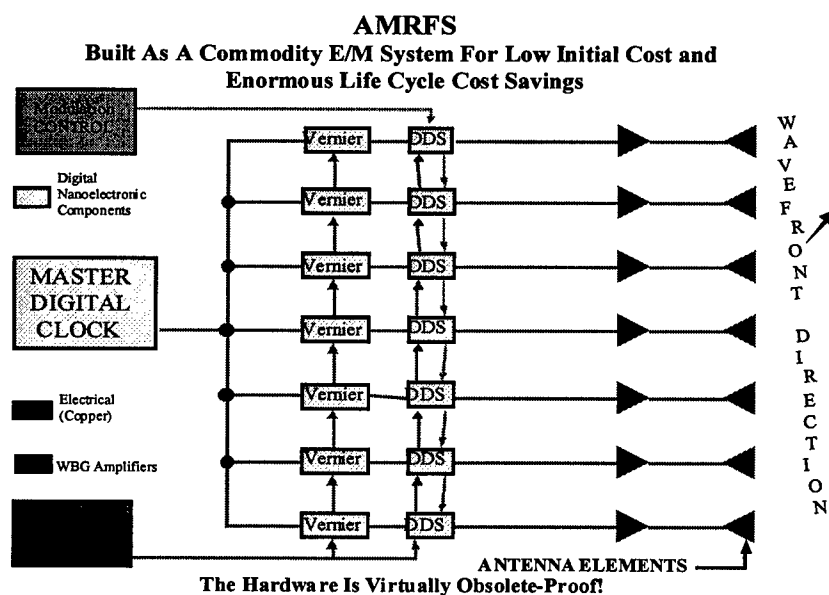


Figure 4. Advanced multifunctional RF system.

among the elements on the edges of an array also precludes precise beam pattern (footprint) control.⁴ The tandem arrangement of MEMS devices can also be used to compensate for this problem.

6. Impact

Versatile, digitally synthesized RF signals with integral MEMS and count down time delay control as well as integral phase and frequency modulation control are expected to provide the basis for an advanced multifunctional rf system (AMRFS) capable of radiating or receiving multiple simultaneous beams of energy. Each of these beams is capable of independent control of azimuth, elevation, power, center frequency, beamshape and information (modulation) content. As such, each signal is capable of independent, real time dynamically assigned functions such as communications, surveillance, target illumination, IFF, weapon control, jamming, deception, or tracking as the environmental scenario dictates. Since the instantaneous bandwidth of the synthesizer is intrinsically capable of decades of spectrum, extremely large time bandwidth product signals may be radiated without beam smear or squint as would be encountered in a conventional phase-shift-steered antenna.

7. Status and conclusions

A program is currently in place to demonstrate large-area LEO of GaN on silicon substrates. Another program will develop the 100 GHz logic circuits necessary to synthesize, modulate, and delay the RF signals. The first results of a 1–5 GHz versatile synthesizer are expected in 2001. A synthesizer capable of synthesizing and modulating signals from audio to 25 GHz is expected in 2003. The companion receiver requires very high performance analog to digital converters and these represent the most challenging aspect of the development. Fortunately, however, it can be shown that the dynamic range required of the A/D converters to meet systems constraints becomes less as the instantaneous bandwidth of the signals increases. Large time bandwidth product signals (incompatible with conventional phase-shift-steered arrays) may actually reduce the dynamic range requirements of the E/M receiver A/D converter by as much as 8 bits.

The basis for an extremely versatile multifunctional rf system is currently within reach.

References

1. A. D. Morrison and Taher Daud, "Low defect, high purity crystalline layers grown by selective deposition," U.S. patent 4,522,661, dated June 11, 1985.
2. O-H. Nam, M. D. Bremser, B. L. Ward, R. J. Nemanich, and R. F. Davis, "Growth of GaN and $\text{Al}_{0.2}\text{Ga}_{0.8}\text{N}$ on patterned substrates via organometallic vapor phase epitaxy," *Jpn. J. Appl. Phys.* **36**, 532L (1997).
3. M. J. W. Rodwell, "500 GHz ultra high speed devices," in: *Proc. IEEE Cornell Conference Advanced Concepts High Speed Semicond. Dev. Circ.*, Ithaca, NY, 1997.
4. H. Steyskal and J. Herd, "Mutual coupling compensation in small array antennas," *IEEE Trans. Antennas Propag.* **38**, 1971 (1990).

Are Coordinated Roadmaps for Compound Semiconductor-Based Technologies Needed? A Proposal for Smarter Investments

Herbert S. Bennett

Semiconductor Electronics Division, National Institute of Standards and Technology, Gaithersburg, MD 20899 U.S.A

Christopher Snowden

Institute of Microwaves and Photonics, University of Leeds, Leeds LS2 9JT, England

Richard Van Atta

Strategy, Forces, and Resources Division, Institute for Defense Analyses, Alexandria, VA 22311 U.S.A

A contribution of the U. S. National Institute of Standards and Technology, not subject to copyright.

1. Introduction

A necessary but not sufficient condition for the success of commercial off-the-shelf (COTS) defense procurements requires improved coordinated planning and priority setting between those in primarily commercial organizations and those in primarily defense organizations. This is especially true for defense systems that contain technologies based on compound semiconductors. In this article we will discuss the need to increase the awareness of decision makers about the challenges associated with COTS procurements and to enhance the consensus-based planning efforts by the defense organizations or establishments. We examine procedures for increasing the probability for successful, knowledge-based investments to support defense procurements. Within the context of these procedures, a proposal to identify compound semiconductor technology challenges in wireless real-time digital video communications networks is used to illustrate that coordinated planning between commercial and defense organizations is worthwhile. We propose the use of existing committees or groups, with rearranged functions in some cases, to build the infrastructure for improved investments in compound semiconductors. These committees would have the following functions:

1. coordinate efforts among those commercial technology roadmaps relevant to defense systems;
2. coordinate among defense organizations their needs for technologies;

3. recommend ways to bridge the technology gaps, where feasible, between what the military needs and what will be available from COTS.

The expected trends in microelectronics for commercial and defense applications suggest that the infrastructure for consensus-based planning needs to be strengthened in both sectors. In general, the resources available to those companies involved with compound semiconductors are much smaller than the resources available to those involved with mainstream silicon complementary metal oxide semiconductors (CMOS). Coordinated planning is one way to invest better the limited resources for compound semiconductor technologies, particularly for defense technologies. For example, since 1993, all 14 U.S. GaAs manufacturers have shifted their emphasis from primarily defense applications to primarily commercial applications.¹ Also, during the last five years some U.S. companies that produce primarily for commercial wireless and microelectronic markets are moving away from compound semiconductors in favor of Si CMOS and BiCMOS.² These shifts place a greater urgency for coordinated, consensus-based plans that should involve both commercial and defense organizations concerned about compound semiconductors. Because of limited resources in any one nation or economic region, the proposed planning for compound semiconductors should be international in scope and involve companies and universities from Asia, Europe, and North America.

Suggesting from the beginning that planning for compound semiconductors be an international effort is consistent with the silicon CMOS efforts at SEMATECH and at the Semiconductor Industry Association (SIA). The formation of a new SEMATECH subsidiary, International SEMATECH, was announced on April 2.³ At present, participation in International SEMATECH is open to current members of the International 300 mm Initiative (I300I). Companies from France, Germany, Korea, The Netherlands, Taiwan, and the United States are members of I300I. Recently, Japanese CMOS manufacturers began their 300 mm initiative called Semiconductor Leading Edge Technologies Inc. (SELETE). Both SEMATECH and the I300I have collaborated with SELETE on international standards for 300 mm silicon wafers. Japanese participation in the new International SEMATECH may be possible in the future. International SEMATECH has programs on lithography infrastructure, standards, and environmental, health and safety. The SIA has decided to internationalize the process for creating the next version of the Technology Roadmap for Semiconductors (TRS).⁴ Organizations from Asia and Europe are being invited to name members for the International Overall Roadmap Technology Characteristics Working Groups. Two members each from Japan, Korea, Taiwan, and the European Community are being sought for each group, to be joined by two members from the U.S. Invitations are going to the Semiconductor Industry Research Institute of Japan (SIRIJ), Korean SIA (KSIA), Electronics Research and Service Organization/Industrial Technology Research Institute (EROS/ITRI) of Taiwan, and either the European Community Association (ECA) or the Micro-Electronics Development for European Applications (MEDEA). The focus of the TRS, an international roadmap, will be on needs of the CMOS industry. Participating organizations will be able to

develop internal domestic versions to address potential solutions as they may desire. Similarly, we propose here that the international planning for compound semiconductors emphasize technical barriers and needs and that participating organizations be free to develop internal domestic plans to provide solutions.

The main purposes of this paper are to:

1. suggest procedures for increasing the awareness of decision makers about the challenges associated with COTS procurements in compound semiconductors;
2. encourage continuous consensus-based planning efforts for successful COTS procurements by defense organizations or establishments; and
3. identify technology challenges in compound semiconductors for a few specific systems that will serve as a focus in planning efforts. It is important that this focus contain enough cross-cutting technologies to be representative of the compound semiconductor manufacturing infrastructure.

Because compound semiconductor technologies are diverse, a focus is needed. Wireless digital video could be a strong candidate for providing this needed focus. Any credible plan for wireless digital video would have to consider compound semiconductor technologies for such applications as front ends of receivers, analog-to-digital converters, and optoelectronic integrated circuits (OEICs).

Some have stated that motivating a compound semiconductor technology roadmap is difficult because, unlike the case for Si CMOS, an equivalent to Moore's Law does not exist and market shares have not been altered sufficiently among competitors to induce one subset of competitors to undertake a technology roadmap. But compound semiconductor field-effect transistors (FETs) and perhaps analog-to-digital converters (ADCs) are in fact constrained by the density of devices. Also, even though many applications for compound semiconductors are not constrained by the increase in the density of devices as a function of time (as described by Moore's Law), equivalents to Moore's Law for such parameters as gain and frequency exist. As interests move above 100 GHz, the transit time impacts on performance, and it can be argued that conventional transit time devices may never break the 1-THz barrier. It is likely that defense organizations will be the first to need solutions to the 1-THz challenges. Many people argue that quantum devices will resolve this issue, but it is not yet clear that this will be the case.

At the other end of the spectrum, one of the key trade-offs between compound semiconductors and Si concerns the availability of appropriate benchmarking standards so that design engineers and manufacturers may communicate with one another to meet production schedules. Some of these benchmarking standards include performance parameters such as gain, efficiency, noise, linearity, power consumption, and thermal management. Most compound semiconductor devices have higher gains with higher operating and maximum frequencies compared to Si devices. But they often have lower thermal conductivities. This difference is often at the hub of the debate on GaAs/AlGaAs heterojunction bipolar transistors (HBTs) versus Si, for example. For many of these quantities, providing credible

graphs of how they vary in time is not straightforward because agreed upon high quality ways to measure these quantities at the highest frequencies are not readily available. This case is one example that demonstrates the requirement in roadmaps for measurement capability. Metrology is one of the crosscut technologies in the National Technology Roadmap for Semiconductors (NTRS). The NTRS and roadmaps from the National Electronics Manufacturing Initiative (NEMI) and the Optoelectronics Industry Development Association (OIDA) identify the role that government has in supporting a metrology infrastructure.

According to some, the market share that GaAs/AlGaAs HBTs now have for 1.8 GHz to 2.5 GHz wireless applications is being challenged by Si BiCMOS technology.^{5,6} The rf applications for BiCMOS in Ref. 5 may become the next example that illustrates how Si CMOS and Si BiCMOS are aggressive technologies that migrate into markets previously dominated by compound semiconductors. Hence, the market share shift that occurred among competing Si CMOS companies and that led after many years to the National Technology Roadmap for Semiconductors (NTRS) is replaced in the case of compound semiconductors with market share losses to Si-based technologies. The application of microelectronic devices to mobile communications from 900 MHz to 6 GHz is a good area for consensus-based planning because it is a mass market driver for which performance and cost are very important, and because it will encompass many Si and compound semiconductor devices.

Even though individual companies and associations have their own compound semiconductor plans, a more comprehensive and globally based plan for one or two applications of compound semiconductors with potentially large markets does not exist. Today, compound semiconductor companies compete on technology, fabrication, and design. This mode of competition among silicon CMOS manufacturers is changing because the research and development costs associated with advanced larger wafer sizes and smaller linewidths for CMOS are too great for any one company or country to accept.⁷ The competitiveness among CMOS manufacturers is shifting from technology and fabrication to product design, supported substantially by advanced computer simulations.

A spectrum for defense procurements of components and systems exists. This spectrum includes:

1. commercial off-the-shelf (COTS);
2. COTS with additional development and some dedicated re-manufacturing (COTS+), and
3. military off-the-shelf (MOTS) with substantial research, development, and dedicated manufacturing.

Smart investments require that technology and material-specific roadmaps (consensus-based plans) be developed by both commercial and defense organizations. Such roadmaps will identify technology gaps and differences between the performance of commercial systems and the performance requirements for derivative defense systems. Once the technology gaps and

performance differences are identified, then defense organizations will have the knowledge base on which to determine funding priorities for research, development, and dedicated manufacturing. Some COTS+ and MOTS always will be necessary in order to maintain the fighting unit's superiority over its challengers. Therefore, funds should be allocated to support an infrastructure for improving defense investments, especially for those technologies that involve wireless and microelectronic digital components and OEICs.

Selected existing organizations in the commercial and government sectors could rearrange or expand their respective activities to include those outlined in the next three paragraphs.⁸ The question of which organizations would participate is answered by the underlying technologies and materials for the systems under consideration. For illustrative purposes to highlight their respective functions, we have given names to three proposed "committees" that need not be new entities and that need not be separate.

These "committees" would have the following functions:

The Committee for Commercial Technology Roadmaps (CCTR) would coordinate efforts among those commercial technology roadmaps that are relevant to defense systems and would identify technologies for which additional roadmaps are needed to plan for expected commercial markets and defense systems. Examples of such commercial roadmaps include NEMI, NTRS, and Semiconductor Equipment and Materials International (SEMI).

The Committee for Performance Specifications of Defense Systems (CPSDS) would coordinate among defense organizations their needs for technologies to support weapons systems to meet their respective missions.

The Advisory Committee for Defense Investments (ACDI) would recommend ways to bridge the technology gaps between what the military needs as determined by the CPSDS and what will be available from COTS as determined by the CCTR and would recommend strategies and priorities for delivering unique defense systems at the cheapest point in the procurement spectrum from COTS to MOTS. This committee would reduce redundancies among components (COTS, COTS+, and MOTS) in weapons systems and would also suggest additional R&D needs beyond COTS for defense programs.

2. Approach — an example

The history of the National Electronics Manufacturing Initiative (NEMI) offers some guidelines on how to proceed with the above proposal.⁹ There are four major lessons learned from NEMI:

1. Have discussions with senior managers in industry to solicit their ideas and support;

2. Hold preliminary workshops to discuss key issues, common interests, and capabilities;
3. Work from a "virtual product" to base materials and vendors; and
4. Select a large enough effort to be effective, but still focused enough to have measurable progress.

In this paper, we are suggesting that wireless real-time digital video (WRTDV) systems be the "virtual product." Digital video in the context of this paper is much broader than its usual meaning. Digital video encompasses very high quality, very high data rates of information (audio, computer, and video) over communication networks. The missions of other government agencies, departments, or ministries would also be served well by WRTDV. Examples include delivery of health care, weather, agriculture, and commerce.

As an illustrative example for carrying out the above proposal and to motivate others to think about our proposal, we consider the objective to have systems that include WRTDV for communications networks. Such systems contain numerous components that span several bands in the electromagnetic wave spectrum; involve microelectronics, optoelectronics, microwave, microelectromechanical (MEMS) and micro-optoelectronics mechanical (MOEMS) systems; require advanced analog-to-digital converters (ADCs), digital-to-analog converters (DACs), digital signal processing (DSP), and optoelectronic integrated circuits (OEICs); incorporate transmitters/sources and receivers/detectors; and depend on unique displays, cameras, and sensors.

Compound semiconductors will be essential for optoelectronics, microwave transmitters and receivers, ADCs, and OEICs. The latter are essential for connecting the wireless sites to fixed and satellite portions of the overall digital video network. The trends and technical challenges for a selected subset of these components are listed in the next two paragraphs to illustrate areas for which international consensus-based planning is appropriate.

Transceiver Front Ends: Today's transceiver front ends are made by mixing and matching transmit/receive switches, low noise and linear amplifiers, power amplifiers, and up/down converters, and other discrete components. Both the commercial and military sectors would benefit by assessments of operating frequency vs. linewidth in FET technology above 10 GHz. The commercial world should begin considering spectrum crowding and the role that frequencies above 10 GHz will play in adding more communications channels. The adoption of high-definition television and enhanced digital services will exacerbate spectrum crowding. The respective roles for HBTs and metal semiconductor field effect transistors (MESFETs) in alleviating spectrum crowding will be determined by economics and technical performance. Technical performance includes such specifications as power efficiency, linearity, circuit complexity, die size, feature size, and reliability (error rates and lifetimes to failure).

ADCs: Digital electronics offers major opportunities for both commercial and military applications. As companies strive to deliver more services to the home, those services will depend on digital electronics. The amount of digital signal

processing in television receivers is increasing dramatically.¹⁰ For both commercial and military applications, getting analog to digital as soon as possible brings economic and technical benefits, particularly for wireless real-time digital video. Most information originates as analog signals. Digital signal processing is much more versatile, cheaper, smaller, and uses less power. Empirical data on the number of bits (resolution) vs. sampling rates show that a boundary exists in the performance of analog-to-digital converters.¹¹ This boundary is moving very slowly towards higher resolutions at the rate of about 1 bit every 8 years. It is conjectured in Ref. 11 that aperture jitter is the main cause of this boundary. Developing the fundamental understanding of the effects that sampling gate timing uncertainties have on ADC performance should be a part of any plan for WRTDV systems. Emerging technologies based on high count InP HBT and high electron mobility transistor (HEMT) circuits and resonant tunneling diodes might move this boundary more quickly.

Some existing organizations that could contribute (with extensions to their present charters and additional resources) to the above proposal on an enhanced infrastructure for wireless digital video for commercial and military applications are suggested below. These suggestions are not complete¹² and serve primarily to present a few specifics about how such an infrastructure might be developed for coordinating commercial and military efforts in WRTDV.

CCTR: Several organizations have already produced technology roadmaps that are relevant to some of the components critical to WRTDV systems. These roadmaps include:

1. Roadmaps from the National Electronics Manufacturing Initiative (NEMI)⁹ that treat numerous market applications and have some material specificity in energy storage devices, radio frequency (rf) devices, and optoelectronic integrated circuits (OEIC) components.
2. Roadmaps from the Optoelectronics Industry Development Association (OIDA)¹³ that treat several diverse market applications and have some material specificity in sensors, detectors, and displays.
3. The National Technology Roadmap for Semiconductors¹⁴ that treats the two very big market applications of microprocessors and memory, that is very material- and process-specific since it is limited to crystalline silicon complementary metal-oxide-semiconductor (Si CMOS), and that has simple metrics for determining progress (e.g., linewidth and density of devices).

CPSDS: Some members of the Microwave Solid-State Electronics Division (MSSSED) of the Electronics Industries Association (EIA) proposed a Microwave Technology Roadmap (MTR) in 1996. The MTR activity was to be composed of an oversight group that would coordinate the deliberations and recommendations of five working groups on 1) design, 2) materials 3) process integration, lithography, devices, and structures, 4) packaging, and 5) test and evaluation. However, the available resources were too limited to develop the MTR.

ACDI: Some countries have groups that provide advice to defense establishments on military applications for microwave, optoelectronics, and microelectronic devices. Such groups could assist in determining costs (R&D, design, manufacturing, testing, maintenance, ownership, and disposal) to meet present and future performance requirements (e.g., frequency/wavelength, bandwidth, power, noise, reliability, and life-cycle costs) of electronic components in WRTDV systems for defense.

3. Issues for wireless digital video networks

We list a few of the many technical COTS issues associated with making WRTDV communications networks a reality. WRTDV involves very high bit rates of data generated, transmitted, processed, stored, and received in a manner so that the synchronization of the digital data stream is maintained at each step. Even though the emphasis here is on video information, other types of high quality data may be carried over of the same network.

1. Commercial specifications and defense specifications for wireless communications components and for materials from which they are made may differ substantially.
2. Commercial products are often available in the marketplace for times that are much less than the lifetime of military systems.
3. The missions of some government agencies may require specifications that limit the applicability of COTS and COTS+. For example, if selected applications for governments to carry out their roles in defense and health care were to require the use of progressive scanning to acquire and display the video, then appropriate COTS equipment for video acquisition by progressive scanning methods may not be available.
4. The military demands on digital video (DV) will most likely be greater than the COTS demands on DV.
5. The requirements on Si CMOS integrated circuits (ICs) for commercial DV¹⁵ are greater and more challenging than some of the performance goals for Si CMOS ICs given in the NTRS.¹⁴ The technical challenges faced by Si CMOS for DV may be summarized by comparing performance parameters given in Figures 2, 5, 7, and 9 in Ref. 15 with the corresponding parameters given in Tables 1 and 2 in Ref. 14.

4. Proposed actions — the next steps

As one of the first steps in building the infrastructure to guide the development of WRTDV, some workshops that identify key subsystems and the technology performance gaps between what is available today and what will be needed to

make WRTDV a reality would be appropriate. These technology gaps should be ranked, if possible, according to the perceived technical difficulty. A few examples, selected here for illustrative purposes, include: 1) miniaturization of microwave filters, 2) low power and very linear microwave amplifiers, 3) circuits with large numbers of heterojunction bipolar transistors (HBTs) for high-speed and high-bit ADCs, 4) increasing the resolution and reducing timing uncertainties in ADCs, 5) long-lived, continuous semiconductor lasers, 6) high resolution displays that are bright and yet consume a minimum of power, and 7) increasing the speed and density of Si CMOS transistors for processing and storing digital video data streams without interruption.

During these proposed workshops, the NEMI model may serve as a useful guide. Industry, academe, and government should agree on selecting a few critical subsystems for WRTDV, and perhaps focus initially on microwave and millimeter wave components and high speed digital electronics such as ADCs and DACs. Then, the participants should identify industry people to champion and to lead and identify government people to gather government procurement data (present and future). Such data may include 1) market sizes in units of products and in dollars, 2) performance specifications likely from COTS and not likely from COTS, and 3) sources and acceptance criteria for the starting materials (Si and GaAs wafers) from which the components would be made.

To develop more convincing arguments that the benefits of consensus-based planning for selected compound semiconductor technologies outweigh the costs associated with planning, the industry/military planning framework should include economic assessments with planning and without planning. Past examples of COTS-dependent procurements should be included. Such examples could include the economic costs of relying on commercial production for displays in airborne command centers and of supporting an industry to re-gun the cathode ray tubes (CRTs) that are no longer available from commercial sources.

Economic assessments of future scenarios will be essential. A challenging assessment will be the one to determine the economic advantages and disadvantages if the military were to select progressive scanning for digital video and the commercial world were to be dominated by interlaced scanning. For example, other assessments could include determining: 1) the savings by not having to maintain life-support systems in reconnaissance airplanes through the use of WRTDV systems; 2) the costs and benefits of microwave front-end technologies based on compound semiconductors vs. those based on silicon CMOS; and 3) the economic advantages and disadvantages of increasing the performance of ADCs through the use of compound semiconductors.

5. Conclusions

We have proposed the use of existing committees, with perhaps rearranged functions in some cases, to build the infrastructure for improved investments in compound semiconductors. These committees would coordinate efforts among

those technology roadmaps relevant to compound semiconductors, coordinate among defense organizations their needs for compound semiconductor technologies, and would recommend ways to bridge the technology gaps, where feasible, between what the military needs and what will be available from COTS. We have suggested that the technology requirements for wireless, real-time digital video in communications networks to meet both commercial markets and government needs could be used to provide a focus among the many diverse compound semiconductor technologies. In those cases for which commercially available products cannot meet the needs for governments to provide such functions as defense, health care, and weather information, alternative methods for determining priorities to support additional research and development have been suggested. The compound semiconductor industry needs improved industry, government, and university collaborations in order for that industry to deliver its full competitive potential.

6. Acknowledgments

One of the authors (H. S. Bennett) thanks Bruce Field, Joseph Pellegrino, and David Seiler for helpful comments and discussions.

References

1. See Table 2 in *Compound Semicond.* 2 (5), 11 (1996); P. T. Greiling, *Compound Semicond.* 2 (2), 52 (1996) and the biennial survey published by the editors of *Microwave J.*, August 1996.
2. "Silicon Update," *Compound Semicond.* 4, 12 (1998).
3. *Technology Business*, p. 7 (March/April 1998); SEMATECH Fact Sheet at <<http://www.sematech.org/member/division/300/home.htm>>.
4. R. I. Scace, private communication, May 1998.
5. Y.F. Chyan, S. Chaudhry, Y. Ma *et al.*, "A very high-performance, epi-free, and manufacturable, 3.3 V 0.35 micrometer BiCMOS technology for wireless applications," in: *Dig. Tech. Papers 1997 Symp. VLSI Technol.*, Kyoto, Japan, June 10-12, 1997, p. 35.
6. R. A. Metzger, *Compound Semicond.* 4 (2), 14 (1998).
7. H. Komiya, "The 300mm technology: current status and future prospect," in *Dig. Tech. Papers 1997 Symp. VLSI Technol.*, Kyoto, Japan, June 10-12, 1997, p. 1.
8. H. S. Bennett, "Summary report on planning for compound semiconductor technology," *J. Res. Natl. Inst. Stand. Technol.* 101, 89 (1996). Copies may be obtained by sending e-mail to the author at herbert.bennett@nist.gov
9. *National Electronics Manufacturing Technology Roadmaps*, National Electronics Manufacturing Initiative, Inc., Herndon, VA, December 1996.

10. A. Thygessen, remarks made during the discussion panel, "PC/TV — TV/PC: where are we headed and who is winning?" 1998 National Association of Broadcasters Program, April 6-9, 1998, Las Vegas, NV, p. 73.
11. R. H. Walden, private communication, May 1998; and "ADC survey and analysis," to be published in the *IEEE J. Selected Areas Commun.*, 1998.
12. Because this report is based on a summary of the authors' discussions, inputs from representatives from the organizations mentioned here have not been solicited. That would be a next step.
13. *Optoelectronic Technology Roadmap*, Optoelectronics Industry Development Association, Washington, D.C., October 1996.
14. *National Technology Roadmap for Semiconductors*, Semiconductor Industry Association, San Jose, CA, 1997 Edition.
15. H. Sasaki, "Multimedia: future and impact for semiconductor technology," *Tech. Digest IEDM* (1997), p. 3.

21st Century: The Final Frontier for III-Nitrides Materials and Devices

Manijeh Razeghi

Center for Quantum Devices, Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60208

1. Introduction

Two of the key components that lie at the heart of information technology are imaging and data storage. Semiconductor materials need to be developed for use in brighter and more colorful light sources for imaging, and in short wavelength light sources for higher density optical data storage. The nitrides of group III elements (or III-nitrides), such as AlN, GaN, InN and their alloys, have all the desired physical properties for this endeavor. They are wide bandgap semiconductors, with a bandgap energy from 6.2 to 1.9 eV, which makes them ideal for optical devices operating in the visible-to-ultraviolet spectral region, such as blue, green, and ultraviolet lasers and photodetectors. Their exceptional physical properties also bring device applications beyond imaging and data storage. These materials are physically and chemically strong, making them ideal for operation in harsh environments (radiation, heat) such as those typical of space applications and high temperature electronics. III-Nitrides also have a high potential for use in high-frequency and high-power electronics. To date, III-nitrides have kept their promises. Super bright green and blue light emitting diodes are just being commercialized. Blue laser diodes have been demonstrated. Experimental electronic devices can operate at temperatures as high as 500 °C. Maximum operating frequencies up to 92 GHz have been demonstrated.

In the next century, the challenges faced by the semiconductor community will be to develop a reliable III-nitride material technology to fulfill the potential of this material system for optoelectronic devices. The effort will have to be articulated around these major issues: nature of the substrate, growth technology, material processing, device structures and modeling, and system applications.

The current lack of a native III-nitride substrate has made the use of foreign substrates inevitable. In the next century, high quality, large area native III-nitride substrates will have to be developed even though this effort may take a long time. In the meantime, novel approaches to avoid defects due to lattice and thermal mismatch between III-nitride films and foreign substrates will have to be developed. An example of such growth technology is lateral epitaxial overgrowth (or LEO). Current device structures will have to be optimized and novel structures developed for new applications. Thanks to improvements in the substrate and growth technology, III-nitride physical properties are expected to be better understood and device modeling more accurate.

This article will first review the physical properties of III-nitride materials, followed by a selective description of the areas of applications. Technological details will then be discussed, including the material growth, the substrates used, and doping. The current state-of-the-art in III-nitride thin films, heterostructures, and optical devices will be presented.

2. Physical properties of III-nitride materials

Unlike more conventional semiconductors, such as silicon (Si) or gallium arsenide (GaAs) which have a diamond or zincblende structure with a cubic symmetry, III-nitride crystals are wurtzite in their stable form with hexagonal symmetry. III-Nitrides are polar crystals and they thus possess many potentially useful properties such as piezoelectricity, pyroelectricity¹ and second harmonic generation.^{2,3} The large difference in electronegativity between the group III and group V elements results in very strong chemical bonds, which are at the origin of most of the interesting III-nitride physical properties.

A direct result of the strong binding is the wide bandgap, ranging from 6.2 to 1.9 eV, corresponding to a wavelength from 200 to 650 nm. This spectral range covers the visible spectrum (blue, green, yellow and red) as well as the near ultraviolet (UV) region in which the atmosphere transmits. The bandgap is direct, which is most appropriate for optical devices. Because the intrinsic carrier concentration is an exponential function of the energy gap and the temperature, a wider bandgap semiconductor has a much lower intrinsic carrier concentration over a large temperature range, resulting in lower leakage and dark currents. Reduced leakage and dark currents are especially important in photodetectors and high temperature electronics.

Another consequence of the strong chemical bonding is the chemical and physical stability (high melting points, mechanical strength, *etc.*) of these materials. They also exhibit high thermal conductivity. Their effective masses are higher than conventional semiconductors, thus leading to lower carrier mobilities, but this drawback is made up for by the high saturated electron drift velocities predicted for this material system. The refractive indices of III-nitrides are lower compared to narrower gap semiconductors, resulting in a lower reflectivity at the interface — an advantage for photodetector efficiency, but a disadvantage when trying to achieve lasers with low threshold currents.

Although III-nitride based devices are being rapidly demonstrated and commercialized, there are many barriers that must be overcome before the full potential of these materials can be realized as reliable devices. First, the high melting points of III-nitrides and the extremely high nitrogen partial pressures near the melting points make their bulk growth very difficult. Therefore, high quality III-nitride substrates do not exist. The synthesis of nitride crystals thus has to be carried out in the form of thin films on a non-native substrate. The dissimilarity between the substrate material and the III-nitrides generally leads to poor structural quality as a result of the lattice and thermal mismatch. Moreover, nitride alloys with different compositions are also lattice-mismatched, which leads to

dislocations in III-nitride heterostructures. Finally, high free electron and hole concentrations are often difficult to achieve because most dopant elements form deep levels in wide bandgap nitride semiconductors. The addition of more dopant source during the growth process frequently results in degradation of the structural and optical properties.⁴⁻⁶

3. Areas of application

The driving force behind the exceptional interest in III-nitride materials has been their potential for numerous significant device applications, both civilian and military. The majority of such devices can be divided into two categories: electronic and optoelectronic (photodetectors and light emitters).

Electronic devices using III-nitrides are needed for both high power and high temperature applications. High power and high temperature devices are in demand by the automotive, aerospace, and power industries. Power electronics have potential applications in the power industry and in any equipment that uses significant power, such as all electric vehicles, ships and aircraft. High temperature electronics would allow control directly in harsh environments, such as in engines, making them important in the automotive and aerospace industries. Because III-nitride devices are expected to better withstand high temperatures while at the same time operate adequately, these electronics could be operated uncooled, thus reducing the cost and weight of systems.

Solar-blind UV photodetectors are sensitive to UV radiation while being (ideally) insensitive to longer wavelength radiation. Such devices have applications where there is a need to detect or control the source of UV radiation in an existing background of visible or infrared radiation.⁷ Examples of such applications include flame detection, furnace and engine monitoring for the automotive, aerospace and petroleum industries, undersea communications, UV astronomy,⁸ space-to-space communications secure from Earth, early missile threat warning and airborne UV countermeasures, and portable battlefield reagent/chemical analysis systems. Because of their theoretical intrinsic solar blindness and low dark currents, III-nitride based devices are expected to work without optical filters and complex electronics, thus significantly reducing the launch weight for space and airborne applications.

III-nitrides have been successfully used in commercial bright blue and green light emitting diodes (LEDs). When used with the already available red AlGaAs based LEDs, these new LEDs complete the primary colors (red, green, blue) for large, high brightness, outdoor full-color displays. Traffic lights are starting to use green LEDs because of their superior efficiency and reliability in comparison to incandescent light sources. Solid state white light sources using a combination of red, green, and blue LEDs or using phosphors excited by blue or ultraviolet (UV) LEDs may soon replace conventional light bulbs with better efficiency and reliability. UV LEDs could also replace the inefficient and hot "black" lights that are used in fun houses, tanning salons, and in more mundane applications such as killing bacteria in water.

The main thrust in recent III-nitride research has been the fabrication of a reliable, short-wavelength (ultraviolet to green spectral region) laser diode. The primary advantage of shorter wavelengths is the ability to focus the beam to smaller spot sizes, which scale as λ^2 , thus quadrupling the storage density of optical media by reducing the laser wavelength in half. The objective will be to achieve a digital video disk (DVD) system capable of storing 15 gigabytes by the year 2000. A DVD-RAM system would require a laser diode operating in continuous wave (CW) at 60 °C with an output power of 30–40 mW and an operating voltage of 3 V at 100 mA. The requirements for a DVD-ROM system would only be a 4–5 mW CW laser. In both cases, the laser should have a wavelength of 400–430 nm. It must not be too short, in order to avoid transmission losses in air. Visible laser diodes are also expected to be used in projective displays, optical communications, and chemical analysis because the wavelength could be tuned to correspond to absorption lines of specific airborne chemicals to be detected. For example, a 55-inch display needs a luminosity of 500 cd/m², which requires at least a 6.6 W red, 1.8 W green, and 1.2 W blue laser. Finally, laser printing is also an important application for short wavelength laser diodes. These would need to emit at a wavelength higher than 430 nm to avoid the decomposition of the toner components, with a single mode CW output power higher than 6 mW for fast printing.

4. Substrates and growth techniques for III-nitrides

Because of their extreme physical properties, bulk III-nitride single crystals are barely available and their quality is not good enough to be used as substrates. Many non-native substrates have been investigated over the years. To date, three substrates stand out as the most promising: silicon, silicon carbide (SiC) and sapphire (Al₂O₃). Silicon is the most widely available substrate in the semiconductor industry and can come in sizes up to 12 inch diameter. It is also the cheapest one and the highest quality. However, it suffers from a poor “compatibility” with III-nitride crystals, meaning a crystal structure (bulk and surface arrangement of atoms) that does not lend itself to the proper initial nucleation of oriented III-nitride crystals. Also, it has a very narrow bandgap in comparison to nitrides, which makes it ill-suited for optical devices. At the other extreme, SiC offers the closest match with III-nitrides in terms of crystal symmetry, lattice and thermal mismatch. The lattice constant is smaller, resulting in compressive strain of the epilayer. On the other hand, the low thermal expansion coefficient of SiC results in tensile strain in the film upon cooling. It is a wide bandgap semiconductor being developed for applications in high power electronics. Its drawbacks are its limited availability, small wafer size, quality still not as good as other Si or Al₂O₃ substrates, and its prohibitively high price.

Sapphire offers a compromise between Si and SiC, and has become the most often used substrate for III-nitride epitaxial growth. The appealing features include the high thermal and chemical stability, the large high quality wafers available, and the reasonable cost. However, there are large lattice and thermal mismatches.

The thermal expansion coefficient is much larger than that of the III-nitrides, resulting in a compressive strain in the epilayer upon cooling. The III-nitrides generally grow on (0001) sapphire substrates with a 30° rotation about the *c*-axis with respect to the sapphire lattice, resulting in the alignment of the directions [00·1]||[00·1] and [10·0]||[11·0]. This orientation results in a compressive strain on the III-nitride layer since the effective lattice constant of the substrate is smaller than those of any III-nitrides: $a_{eff} = a/\sqrt{3} = 2.747 \text{ \AA}$. The dissimilarity between the substrate and nitride materials has been alleviated through the successful development of the growth technology, and more precisely the use of buffer layers.⁹ As it seems that a reliable source of III-nitride substrates is still far away, the substrate of the near future for III-nitride devices will most likely be sapphire.

One of the reasons that III-nitrides suffered from a lack of interest until the 1980s was the lack of suitable growth techniques. Currently metalorganic chemical vapor deposition (MOCVD) and molecular beam epitaxy (MBE) are the most widely used techniques. Recently, vapor phase epitaxy (VPE) growth of III-nitrides has gained attention for the growth of very thick GaN films for use as substrates after original substrate lift-off.¹⁰ High-pressure growth remains limited to a few research groups in Poland for the direct growth of bulk GaN crystals. MOCVD will undoubtedly be the method of choice of the future for the growth of high quality III-nitrides films for mass production and particularly for devices.

5. State of the art of III-nitride thin films

AlN thin films are generally grown on basal plane Al₂O₃ or SiC substrates, without a buffer layer, due to the fact that sapphire and AlN share a common element, aluminum, which makes the bonding at the interface much easier. Epitaxial films are rarely thicker than 1–1.5 μm. High crystalline quality films have been achieved on Al₂O₃ and SiC substrates with open-detector symmetric x-ray rocking-curve linewidths as low as 90 and 60 arc-seconds respectively.^{11,12}

As grown AlN films are almost always insulating. Negative electron affinity (NEA) has also been reported from AlN films.¹³ This effect has been used to demonstrate cold cathode structures using AlN films.¹⁴ The piezoelectric properties of AlN have been investigated for a number of years for surface acoustic wave (SAW) applications.¹⁵ The optical properties of AlN are assessed through optical absorption and cathodoluminescence. Second harmonic generation from AlN films has been conducted and yielded the nonlinear coefficients $\chi^{(2)}_{zzz} = 10 \text{ pm/V}$ and $\chi^{(2)}_{zxx} = 0.5 \text{ pm/V}$. Future research work on AlN includes the development of a bulk substrate, further studying its potential use as a dielectric in III-nitride based electronics, as well as the possibility to achieve low resistivity *n*-type and *p*-type AlN. The nonlinear optical properties of AlN are also of interest.

GaN is by far the most studied III-nitride material. A thin AlN, GaN, or AlGaN buffer layer is generally used for the growth. Basal plane Al₂O₃ and SiC substrates are most commonly used. Films as thick as 100 μm have been reported, depending on the growth technique utilized. High crystalline quality GaN thin films have been achieved, with open-detector x-ray rocking-curve

linewidths as low as 30 arc-seconds and asymmetric x-ray rocking curve linewidths as low as 400 arc-seconds.¹² Undoped GaN films are usually either highly resistive or exhibiting *n*-type conduction with a residual carrier concentration $\sim 10^{16} \text{ cm}^{-3}$ at room temperature and an electron mobility as high as $900 \text{ cm}^2/\text{V}\cdot\text{s}$ (theoretical calculations show that the maximum 300 K electron mobility in GaN is $2350 \text{ cm}^2/\text{V}\cdot\text{s}$). The pyroelectricity properties of GaN have been measured and yield a pyroelectric voltage coefficient of $\sim 10^4 \text{ V/m}\cdot\text{K}$.

The optical properties of GaN are usually assessed through optical transmission and photoluminescence (PL). Free excitons A, B and C have been observed with peak linewidths of $\sim 1\text{--}3 \text{ meV}$ at 2 K using photoluminescence (PL). The room temperature PL linewidths are typically as low as $\sim 30 \text{ meV}$. Residual acceptor and donor related luminescence transitions are often observed as well. A broad "yellow" luminescence is sometimes observed and has been attributed to defects in GaN.

Ternary $\text{Al}_x\text{Ga}_{1-x}\text{N}$ has been grown over the entire compositional range. The resistivity of $\text{Al}_x\text{Ga}_{1-x}\text{N}$ was found to increase exponentially with Al concentration.¹⁶ Low Al concentration alloys sometimes show limited *n*-type conduction due to residual donors as in the case of GaN. The optical properties of $\text{Al}_x\text{Ga}_{1-x}\text{N}$ have been assessed using cathodoluminescence and optical absorption, in particular to determine the bandgap energy. Future research work on $\text{Al}_x\text{Ga}_{1-x}\text{N}$ alloys will be to understand their electrical properties as a function of Al concentration. This will in turn help the *n*-type and *p*-type doping in the entire alloy composition range, thus broadening the range and performance of III-nitride based devices.

One of the main advantages of the III-nitrides over other wide bandgap materials such as SiC is the potential to fabricate heterostructures and achieve bandgap engineering within the same material system. However, AlGaIn/GaN heterostructures are still in their infancy and much remains to be done. A two dimensional electron gas (2DEG) has been demonstrated at the AlGaIn/GaN interface. Room temperature electron mobilities as high as $2000 \text{ cm}^2/\text{V}\cdot\text{s}$ have been measured for a sheet carrier density of 10^{13} cm^{-2} . At 20 K, the electron mobilities were as high as $5700 \text{ cm}^2/\text{V}\cdot\text{s}$ and $7500 \text{ cm}^2/\text{V}\cdot\text{s}$ for structures on sapphire and SiC substrates for a sheet carrier density of $5 \times 10^{12} \text{ cm}^{-2}$.¹⁷

Band alignments in the III-nitride material system have been investigated and remain a controversial issue. Theoretical calculations estimated the valence band offsets of (wurtzite) AlN, GaN, and InN to be : $\text{AlN/GaN} = 0.7\text{--}0.81 \text{ eV}$ and $\text{GaN/InN} = 0.3\text{--}0.48 \text{ eV}$. Experimental measurements of the valence band offsets yielded: $\text{AlN/GaN} = 0.70\text{--}1.36 \text{ eV}$, $\text{GaN/InN} = 1.05 \text{ eV}$ and $\text{AlN/InN} = 1.81 \text{ eV}$.¹⁸

Ternary $\text{Ga}_{1-x}\text{In}_x\text{N}$ has been grown over the entire composition range. However, the material quality significantly deteriorates as the In concentration increases. It was shown that the $\text{Ga}_{1-x}\text{In}_x\text{N}$ alloy composition and material quality very strongly depended on the growth conditions, in particular the growth temperature, growth pressure, V/III ratio, and growth rate. To grow alloys with higher In concentration, it is generally necessary to lower the growth temperature, raise the growth pressure, and increase the V/III ratio and the growth rate.¹⁹

Moreover, it has been reported that GaN and InN have a miscibility gap.²⁰ Ga_{1-x}In_xN films are generally thin (< 0.5 μm) and are grown on thick GaN films (> 1 μm) on basal plane Al₂O₃ or SiC substrates. The x-ray rocking curve linewidths can be as low as 480 arc-seconds (for 14% In).²¹

As grown Ga_{1-x}In_xN films generally show *n*-type conduction ($n > 10^{17} \text{ cm}^{-3}$ at 300 K). Room temperature photoluminescence measurements showed that Ga_{1-x}In_xN can have a linewidth as low as 70 meV (for 14% In). Ellipsometry has yielded the refractive index of Ga_{1-x}In_xN to be ~0.05 higher (for $x = 0.06$) than that of GaN. Future research work on Ga_{1-x}In_xN compounds includes improving their structural and optical quality for $x > 0.5$, as well as their uniformity and homogeneity over large area wafers.

GaNN/GaN heterostructures and quantum wells have been reported, although they have been more often characterized in actual devices.^{22,23} The cathodoluminescence intensity was shown to increase by several orders for GaInN/GaN multi-quantum wells compared to bulk GaInN films.²⁴ There have been reports of a "composition pulling effect" in thin GaInN films grown on GaN.²⁵ It was shown that the lattice mismatch between the growing GaInN layer and the underlying GaN prevented the incorporation of indium into the lattice. This effect can be significant in the control of the emission wavelength from GaInN/GaN quantum wells. Finally, the quantum confined Stark effect in GaInN/GaN multiquantum wells due to piezoelectric effects has been recently reported to influence the optical properties of these structures.²⁶ This effect can be minimized by adequately doping the structures with Si. The strong impetus to fabricate and produce visible optical devices has left much of the fundamental research in Ga_{1-x}In_xN/GaN heterostructures for the future. Although such study is necessary, it will strongly depend on the improvement of the Ga_{1-x}In_xN material quality.

6. Doping of III-nitrides

In order to fabricate devices, it is necessary to control the doping of III-nitrides. The *n*-type doping in these materials has generally been much easier than the *p*-type doping, mainly because III-nitrides have the tendency to exhibit *n*-type conductivity as grown. Like other III-V semiconductors, the *n*-type doping can be achieved using group VI elements, while the *p*-type doping is achieved by incorporating group II elements. Group IV elements, such as Si and Ge, act as donors in III-nitrides, whereas C seems to act as an acceptor. Si and Ge have an electronegativity closer to Al, Ga, and In than N, and thus would be more likely to replace Al, Ga, and In than N. The electronegativity of C is closer to N than to the group III elements, and thus it would be more likely to replace N in the III-nitride lattice.

The *n*-type doping of GaN films has been investigated using Si, Ge,²⁷ Se, S²⁸ and O. The most successful dopants have been Si and Ge. Doping control has been achieved up to a carrier concentration of 10^{20} cm^{-3} . The Si level in the bandgap was estimated to be $\sim 22 \pm 4 \text{ meV}$ below the bottom of the conduction band.²⁹ Impurity band conduction is usually observed at low temperatures.

The *n*-type doping of $\text{Al}_x\text{Ga}_{1-x}\text{N}$ has been achieved using Si and Ge for Al content $x < 0.6$.³⁰ More research is necessary into the *n*-type doping of $\text{Al}_x\text{Ga}_{1-x}\text{N}$ for $x > 0.6$ and of other III-nitrides. This capability would allow better optical and electrical confinement in heterostructures, as well as fulfillment of the potential of III-nitrides for ultraviolet optoelectronics. In order to achieve this, it may prove necessary to understand the origin of the high resistivity of $\text{Al}_x\text{Ga}_{1-x}\text{N}$ for high Al concentrations.

The *p*-type doping of GaN, $\text{Al}_x\text{Ga}_{1-x}\text{N}$, and $\text{Ga}_{1-x}\text{In}_x\text{N}$ films has been achieved using Mg. The doping control is not easy at all as it is very sensitive to dopant flow rate. As-doped films are generally insulating (except a few reports of as-grown *p*-type GaN by MBE) and require post-growth treatment such as thermal annealing ($> 600^\circ\text{C}$ under nitrogen or vacuum) or low energy electron beam irradiation (LEEBI) to activate the *p*-type dopant.^{31,32} The mechanism by which this activation happens has been identified as the breaking of Mg-H bonds.³³ The concentration of Mg atoms in the lattice is typically $< 10^{19}\text{ cm}^{-3}$, but the room temperature free hole concentrations are generally $< 5 \times 10^{18}\text{ cm}^{-3}$ for a mobility $< 20\text{ cm}^2/\text{V}\cdot\text{s}$. The activation energy of Mg has been estimated to be 150–200 meV. Impurity band conduction is also observed at low temperatures in *p*-type GaN films.

The *p*-type doping of $\text{Al}_x\text{Ga}_{1-x}\text{N}$ and $\text{Ga}_{1-x}\text{In}_x\text{N}$ has been carried out using Mg for Al content $x < 0.3$ and In content $x < 0.09$, respectively.^{17,34} Much more research work on the *p*-type doping of III-nitrides is necessary in the future. In particular, a significant increase in the *p*-type conductivity of GaN and $\text{Al}_x\text{Ga}_{1-x}\text{N}$ for $x > 0.3$ is strongly desired. Research directions include new doping sources and new doping schemes involving co-doping.

7. Optical devices

The development of photodetectors based on III-nitride material began with the simplest device — the photoconductor. This device requires no *p*-GaN layer to operate, and therefore simplifies not only the growth demands, but also the fabrication steps because etching steps are not needed to define the device or contact two electrically different types of GaN layers. The characteristics of current photoconductor devices include very fast response,³⁵ $\text{Al}_x\text{Ga}_{1-x}\text{N}$ detectors over the entire range ($0 \leq x \leq 1$),³⁶ and demonstration of very high gain in an interdigitated MSM detector with a responsivity R of over 3000 A/W.³⁷ GaN detector arrays have also been demonstrated. The kinetics of photoconductivity have also been investigated in GaN photodetectors.³⁸

Schottky diodes are the simplest detectors to fabricate and are capable of being extremely fast. Several groups have combined efforts to develop Schottky photodetectors with high responsivity R and low noise equivalent power (NEP).³⁹ To date, *pn* junction photodiodes have all been formed using GaN, leading to a cut-off wavelength of 365 nm corresponding to the bandgap of GaN. The temporal response for these detector designs led to carrier lifetimes as low as 20 ns, with

Group	Detector	R (A/W)	NEP (W)	Lifetime	RR
Boston Univ.	GaN PC	125		20 ns	
APA Optics	$\text{Al}_x\text{Ga}_{1-x}\text{N}$ PC	18–300	$< 10^{-8}$	1–2 ms	10^3
CQD-Northwestern	$\text{Al}_x\text{Ga}_{1-x}\text{N}$ PC	> 10	$< 10^{-9}$	< 0.3 ms	10^3
Univ. Texas–Austin	GaN MSM	0.3			$> 10^2$
NASA-Goddard	GaN MSM	3200		300 μs	$> 10^2$
APA–Texas Tech	n -GaN Schottky	0.18	5×10^{-9}	120 ns	10
APA–NC State	n -GaN Schottky	0.10	4×10^{-9}		10^2
SVT–U. Minnesota	GaN p - i - n	0.11		8.2 μs	
APA–Texas Tech	GaN p - π - n	0.10	4×10^{-11}	18 ns	10^3
Illinois–Wright Lab	GaN/AlGaN p - i - n	0.14	8×10^{-12}	12 ns	10^3
CQD-Northwestern	GaN p - i - n	0.15		2.5 μs	10^6

Table 1. Major accomplishments in nitride-based detectors: photoconductors, MSMs, Schottky diodes, and photovoltaic detectors (see Refs. 35–39 and therein).

very low NEP, good responsivities (0.10–0.14 A/W) and rejection ratios (RR) of about three orders of magnitude. The progression to faster response, higher rejection of visible light,⁴⁰ lower noise interference, and better response are all indicative of not only improved designs, but also a continued increase in the quality of material. Table 1 summarizes some of the nitride-based photoconductor and photodetector results reported in recent years. III-Nitride based UV photodetectors remain a very promising field for research and development for the future.

To generate visible light using III-nitride materials, one has to use $\text{Ga}_{1-x}\text{In}_x\text{N}$ alloys in the active layer. The first generation of blue and blue/green LEDs were fabricated from GaInN/AlGaN double-heterostructures (DH).⁴¹ Although these provided high optical output, higher than 1 candela (cd), they had a broad spectrum with linewidths typically ~ 70 nm, while the emission spectrum ranged from the violet to the yellow-orange spectral range, making the output appear "whitish-blue" to the human eye. Greatly improved LED performance, in terms of both color purity and intensity, has been achieved using single quantum-well (SQW) structures.^{42–44} The emission peak linewidths for blue LEDs (450 nm) have been reduced to 20 nm, with brightness as high as 2 cd. Green LEDs (520 nm) exhibited a emission peak linewidth of 30 nm and luminous intensity of 12 cd. The latest record has been achieved by using strained single quantum wells of $\text{Ga}_{1-x}\text{In}_x\text{N}$. Violet (405 nm), blue (450 nm) and green (520 nm) LEDs have been demonstrated with efficiencies of 9.2%, 8.7% and 6.3% respectively. These LEDs perform better than those made from other materials and join the mainstream of

LED evolution. White LEDs have also been demonstrated by combining a blue nitride LED with a phosphorescent coating of the LED packaging. All these LEDs are now commercially available.

Despite the successful commercialization of III-nitride based LEDs, there remain some important issues. First is the reliability of the devices. These LEDs are still fragile and require careful handling. They can be easily damaged by reverse bias greater than 5 V or forward current higher than 100 mA. This vulnerability is surprising because III-nitrides are expected to have high breakdown voltages. An improper doping profile or high background carrier concentration in GaInN could be the cause of failure at high reverse bias or large forward current. The second issue is their thermal handling capability. The recommended operating temperature is $< 80^\circ\text{C}$ to avoid early degradation of the devices. Ideally, III-nitride based devices are suitable for much higher temperature operation because of their thermal properties. The unusually low operating temperature could be due to the high defect density in the materials and the thermal instability of the GaInN active layer. The third issue is the price. Commercial III-nitride based LEDs cost more than ~\$6 (U.S.) apiece, which is still too high for large volume applications.

The realization of III-nitride based laser diodes eluded the research community for many years. It has been only after thorough development of the material, processing, and device fabrication technology that such lasers have been made possible. There has been outstanding success in GaN-based blue laser diodes in the past few years. To date, nine research groups worldwide have demonstrated GaInN/GaN multiple quantum well based violet-blue laser diodes. Some of the critical device parameters are summarized in Table 2. Blue laser diodes operating

Group	Type	Threshold	$\lambda/\Delta\lambda$ (nm)	Power	Lifetime
Nichia	MQW	4 kA/cm ² (34 V)	417/1.6	215 mW	(> 24 hrs)
Mejo Univ.	SQW	2.9 kA/cm ² (16 V)	376/0.15		
Toshiba	MQW	50 kA/cm ² (20 V)	417.5/0.15		
Cree	??				
Fujitsu	MQW	12 kA/cm ² (22 V)	405-425	80 mW	
UCSB	MQW	12.7 kA/cm ² (50 V)	420		
Xerox	MQW	25 kA/cm ²	422-432		
Sony	MQW	9.5 kA/cm ²	417.5		
Northwestern	MQW	1.4 kA/cm ² (77 K)	405-410	2 mW	140 hrs
Nichia	MQW	1.5 kA/cm ²	390-420	2 mW	10,000 hrs

Table 2. Reported GaN-based laser results as of spring, 1998.

at room temperature and in continuous wave mode with a projected lifetime of 20,000 hours have been demonstrated and are nearing commercialization. Most of the recent nitride lasers share the following characteristics:

- The active layer consists of a $\text{Ga}_{1-x}\text{In}_x\text{N}/\text{Ga}_{1-y}\text{In}_y\text{N}$ or GaInN/GaN multi-quantum well (MQW) with an emission peak ranging from 400 to 430 nm.
- The MQW is not uniform, but has a quantum-dot-like structure. These quantum dots are formed most likely because of indium composition fluctuation and segregation.
- Laser performance is enhanced by Si doping of the wells and barriers of the MQW.
- The general structure of the laser consists of a separate confinement heterostructure, using AlGaIn as the cladding layers.
- The p -type contact layer has a hole concentration in the range of 10^{18} cm^{-3} .
- The dislocation density in the early generation of III-nitride based lasers was measured to be higher than 10^7 cm^{-2} , which led to very short lifetimes and low output power. It is only recently that much higher output power and longer lifetimes have been achieved by reducing the dislocation density through lateral epitaxial overgrowth (LEO).

From the above discussion, it can be seen that there are still many fundamental issues that need to be addressed. First, there is currently no real understanding of the lasing mechanism in III-nitride lasers. There is experimental evidence that recombination in the MQW active layer is enhanced in these laser diodes by self-formed quantum-dot-like structures, or by localization of excitons by potential fluctuation.⁴⁵⁻⁴⁹ Theoretical work is necessary to study how these structures are affecting material and modal gain,⁵⁰ recombination efficiency, emission wavelength (tunable by adjusting the dot or potential feature sizes), and how lasing can be improved by intentionally controlling the formation of such structures. Secondly, the mechanism for the formation of the aforementioned quantum-dot structures or local potentials is also unknown. Is it due to the intrinsic nature of GaInN ternary alloys since compositional modulation due to phase separation would be energetically favored in this material system?²⁰ Or is it due to compressive strain introduced by lattice mismatch? The understanding of the mechanism would inevitably lead to better devices.

The effect of doping in III-nitride based lasers is not entirely clear. Generally lasers have intrinsic active layers in order to enhance carrier diffusion and reduce free-carrier absorption in the active layer. However, Si doping in III-nitride lasers enhances the laser diode performance.⁴⁸ It is believed that the doping effectively screens the piezoelectric field in the MQW active region.⁵¹ The p -type doping needs to be increased in order to minimize device resistance. Searching for other dopants seems hopeless because of the deep-level nature of acceptor dopants in III-nitrides, which may be an intrinsic nature of wide bandgap nitride materials, just like ZnSe -based materials. However, new doping schemes such as the

piezoelectric effect, which can enhance the ionization of impurities by a built-in electric field, or tunneling-assisted carrier injection should be studied.

Finally, the failure mechanisms in III-nitride lasers need to be determined and minimized. Potential causes of failure include heat generation due to high series resistance, dislocations and other threading defects, as well as optical damage due to reabsorption of stimulated emission at defects that are formed during growth.⁵² Reduction of defect density is now rapidly being conducted through lateral epitaxy overgrowth.⁵³ With the progress of bulk GaN growth either by the high-pressure technique, or hydride VPE, or LEO, homoepitaxy of III-nitride devices may someday be available.

8. Conclusion

This review has shown that indeed much has already been achieved in the III-nitride material system over the past two decades, in particular the realization of commercial optical devices. Much more fundamental work still needs to be conducted in order to understand the very rich and unexplored physical properties of III-nitrides. A better understanding of all the physical parameters would allow, in turn, a better design of existing devices and the potential discovery of novel devices.

References

1. A. D. Bykhovski, V. V. Kaminski, M. S. Shur, Q. C. Chen, and M. A. Khan, "Pyroelectricity in gallium nitride thin films," *Appl. Phys. Lett.* **69**, 3254 (1996).
2. J. Miragliotta, D. K. Wickenden, T. J. Kistenmacher, and W. A. Bryden, "Linear and nonlinear optical properties of GaN thin films," *J. Opt. Soc. Am. B* **10**, 1447 (1993).
3. H. Y. Zhang, X. H. He, Y. H. Shih, *et al.*, "Study of nonlinear optical effects in GaN:Mg epitaxial film," *Appl. Phys. Lett.* **69**, 2953 (1996).
4. O. Madelung, ed., *Semiconductors: group IV elements and III-V compounds*, Berlin: Springer-Verlag, 1991.
5. O. Madelung, *Semiconductors – Basic Data*, 2nd ed., Berlin: Springer-Verlag, 1996.
6. M. Leszczynski, T. Suski, P. Perlin, *et al.*, "Lattice constants, thermal expansion and compressibility of gallium nitride," *J. Phys. D* **28**, A149 (1995).
7. M. Razeghi and A. Rogalski, "Semiconductor ultraviolet detectors," *J. Appl. Phys.* **79**, 7433 (1996).
8. P. Kung, A. Saxler, X. Zhang, D. Walker, and M. Razeghi, "GaN, GaAlN, and AlN for use in UV detectors for astrophysics: an update," in: Gail J. Brown and Manijeh Razeghi, eds., *Photodetectors: Materials and Devices*, SPIE Proceedings Series Vol. 2685, Bellingham, WA: SPIE, 1996, p. 126.

9. H. Amano, N. Sawaki, I. Akasaki, and Y. Toyoda, "Metalorganic vapor phase epitaxial growth of a high quality GaN film using an AlN bufferlayer," *Appl. Phys. Lett.* **48**, 353 (1986).
10. L. T. Romano, B. S. Krusor, and R. J. Molnar, "Structure of GaN films grown by hydride vapor phase epitaxy," *Appl. Phys. Lett.* **71**, 2283 (1997).
11. A. Saxler, P. Kung, C. J. Sun, E. Bigan, and M. Razeghi, "High quality aluminum nitride epitaxial layers grown on sapphire substrates," *Appl. Phys. Lett.* **64**, 339 (1994).
12. P. Kung, A. Saxler, X. Zhang, *et al.*, "High quality AlN and GaN epilayers grown on (00·1) sapphire, (100) and (111) silicon substrates," *Appl. Phys. Lett.* **66**, 2958 (1995).
13. M. C. Benjamin, C. Wang, R. F. Davis, and R. J. Nemanich, "Observation of a negative electron affinity for heteroepitaxial AlN on (6H)-SiC(0001)," *Appl. Phys. Lett.* **64**, 3288 (1994).
14. A. T. Sowers, J. A. Christman, M. D. Bremser, *et al.*, "Thin films of aluminum nitride and aluminum gallium nitride for cold cathode applications," *Appl. Phys. Lett.* **71**, 2289 (1997).
15. K. Kaya, H. Takahashi, Y. Shibata, Y. Kanno, and T. Hirai, "Experimental surface acoustic wave properties of AlN thin films on sapphire substrates," *Jpn. J. Appl. Phys.* **36**, 307 (1997).
16. P. Kung, A. Saxler, D. Walker, *et al.*, " $\text{Al}_x\text{Ga}_{1-x}\text{N}$ -based materials and Heterostructures," *Mater. Res. Soc. Proc.* **449**, 79 (1997).
17. J. M. Redwing, M. A. Tischler, J. S. Flynn, *et al.*, "Two-dimensional electron gas properties of AlGa_N/GaN heterostructures grown on 6H-SiC and sapphire substrates," *Appl. Phys. Lett.* **69**, 963 (1996).
18. G. Martin, A. Botchkarev, A. Rockett, and H. Morkoç, "Valence-band discontinuities of wurtzite GaN, AlN, and InN heterojunctions measured by x-ray photoemission spectroscopy," *Appl. Phys. Lett.* **68**, 2541 (1996).
19. A. Koukitu, N. Takahashi, T. Taki, and H. Seki, "Thermodynamic analysis of the MOVPE growth of $\text{In}_x\text{Ga}_{1-x}\text{N}$," *J. Crystal Growth* **170**, 306 (1997).
20. I.-H. Ho and G. B. Stringfellow, "Solid phase immiscibility in GaInN," *Appl. Phys. Lett.* **69**, 2701 (1996).
21. S. Nakamura and T. Mukai, "High-quality InGa_N films grown on GaN films," *Jpn. J. Appl. Phys.* **31**, L1457 (1992).
22. A. Sohmer, J. Off, H. Bolay, *et al.*, "GaInN/GaN-heterostructures and quantum wells grown by metalorganic vapor-phase epitaxy," *MRS Internet J. Nitride Semicond. Res.* **2**, 14 (1997).
23. P. Kung, A. Saxler, D. Walker, *et al.*, "GaInN/GaN multi-quantum well laser diodes grown by low-pressure MOCVD," *MRS Internet J. Nitride Semiconductor Res.* **3**, 1 (1998).
24. M. Koike, S. Yamasaki, S. Nagai, *et al.*, "High-quality GaInN/GaN multiple quantum wells," *Appl. Phys. Lett.* **68**, 1403 (1996).
25. K. Hiramatsu, Y. Kawaguchi, M. Shimizu, *et al.*, "The composition pulling effect in MOVPE-grown InGa_N on GaN and AlGa_N and its TEM characterization," *MRS Internet J. Nitride Semicond. Res.* **2**, 6 (1997).

26. T. Deguchi, A. Shikanai, K. Torii, *et al.*, "Luminescence spectra from InGaN multiquantum wells heavily doped with Si," *Appl. Phys. Lett.* **72**, 3329 (1998).
27. X. Zhang, P. Kung, A. Saxler, D. Walker, and M. Razeghi, "Observation of room temperature surface-emitting stimulated emission from GaN:Ge by optical pumping," *J. Appl. Phys.* **80**, 6544 (1996).
28. A. Saxler, P. Kung, X. Zhang, *et al.*, "GaN doped with sulfur," in: Gordon Davies and Maria Helena Nazaré, eds., *Defects in Semiconductors ICDS-19*, Materials Science Forum Vols. 258-263, Switzerland: Trans Tech Publications, 1998, p. 1161.
29. W. Gotz, N. M. Johnson, C. Chen, *et al.*, "Activation energies of Si donors in GaN," *Appl. Phys. Lett.* **68**, 3144 (1996).
30. X. Zhang, P. Kung, A. Saxler, *et al.*, "Growth of $\text{Al}_x\text{Ga}_{1-x}\text{N}:\text{Ge}$ on sapphire and silicon substrates," *Appl. Phys. Lett.* **67**, 1745 (1995).
31. S. Nakamura, T. Mukai, M. Senoh, and N. Iwasa, "Thermal annealing effects on *p*-type Mg-doped GaN films," *Jpn. J. Appl. Phys.* **31**, L139 (1992).
32. H. Amano, M. Kito, K. Hiramatsu, and I. Akasaki, "P-type conduction in Mg-doped GaN treated with low-energy electron beam irradiation (LEEBI)," *Jpn. J. Appl. Phys.* **28**, L2112 (1989).
33. S. Nakamura, N. Iwasa, M. Senoh, and T. Mukai, "Hole compensation mechanism of *p*-type GaN films," *Jpn. J. Appl. Phys.* **31**, 1258 (1992).
34. S. Yamasaki, S. Asami, N. Shibata, *et al.*, "P-type conduction in Mg-doped $\text{Ga}_{0.91}\text{In}_{0.09}\text{N}$ grown by MOCVD," *Appl. Phys. Lett.* **66**, 1112 (1995).
35. J. C. Carrano, P. A. Grudowski, C. J. Eiting, *et al.*, "Very low dark current metal-semiconductor-metal ultraviolet photodetectors fabricated on single-crystal GaN epitaxial layers," *Appl. Phys. Lett.* **70**, 1992 (1997).
36. D. Walker, X. Zhang, P. Kung, *et al.*, "AlGaN ultraviolet photoconductors grown on sapphire," *Appl. Phys. Lett.* **68**, 2100 (1996).
37. Z. C. Huang, D. B. Mott, P. K. Shu, J. C. Chen, and D. K. Wickenden, "Improvements of metal-semiconductor-metal GaN photoconductors," *J. Electronic Mater.* **26**, 330 (1997).
38. P. Kung, X. Zhang, D. Walker, *et al.*, "Kinetics of photoconductivity in *n*-type GaN photodetector," *Appl. Phys. Lett.* **67**, 3792 (1995).
39. Q. Chen, J. W. Yang, A. Osinsky, *et al.*, "Schottky barrier detectors on GaN for visible-blind ultraviolet detection," *Appl. Phys. Lett.* **70**, 2277 (1997).
40. D. Walker, A. Saxler, P. Kung, *et al.*, "Visible-blind GaN *p-i-n* photo-diodes," *Appl. Phys. Lett.* **72**, 3303 (1998).
41. A. Saxler, K. S. Kim, D. Walker, *et al.*, "Electroluminescence of III-Nitride double heterostructure light emitting diodes with silicon and magnesium doped InGaN," in Gordon Davies and Maria Helena Nazaré, eds., *Defects in Semiconductors ICDS-19*, Materials Science Forum Vols. 258-263, Switzerland: Trans Tech Publications, 1998, p. 1229.
42. S. Nakamura, M. Senoh, N. Iwasa, and S. Nagahama, "High-power InGaN single-quantum-well-structure blue and violet light-emitting diodes," *Appl. Phys. Lett.* **67**, 1868 (1995).

43. S. Nakamura, M. Senoh, N. Iwasa, and S. Nagahama, "High-brightness InGaN blue, green and yellow LEDs with quantum well structures," *Jpn. J. Appl. Phys.* **34**, L797 (1995).
44. S. Nakamura, M. Senoh, N. Iwasa, and S. Nagahama, "Superbright green InGaN single-quantum-well-structure LEDs," *Jpn. J. Appl. Phys.* **34**, L1332 (1995).
45. S. Nakamura, M. Senoh, S. Nagahama, *et al.*, "Longitudinal mode spectra and ultrashort pulse generation of InGaN multiquantum well structure laser diodes," *Appl. Phys. Lett.* **70**, 616 (1996).
46. S. Nakamura, M. Senoh, S. Nagahama, *et al.*, "Room-temperature continuous-wave operation of InGaN multi-quantum-well structure laser diodes with a lifetime of 27 hours," *Appl. Phys. Lett.* **70**, 1417 (1996).
47. S. Nakamura, M. Senoh, S. Nagahama, *et al.*, "Subband emissions of InGaN multi-quantum-well laser diodes under room-temperature continuous wave operation," *Appl. Phys. Lett.* **70**, 2753 (1997).
48. S. Chichibu, K. Wada, and S. Nakamura, "Spatially resolved cathodoluminescence spectra of InGaN quantum wells," *Appl. Phys. Lett.* **71**, 2346 (1997).
49. Y. Narukawa, Y. Kawakami, M. Funato, *et al.*, "Role of self-formed InGaN quantum dots for exciton localization in the purple laser diode emitting at 420 nm," *Appl. Phys. Lett.* **70**, 981 (1997).
50. Y.-K. Song, M. Kuball, A. V. Nurmikko, *et al.*, "Gain characteristics of InGaN/GaN quantum well diode lasers," *Appl. Phys. Lett.* **72**, 1418 (1998).
51. T. Deguchi, A. Shikanai, K. Torii, *et al.*, "Luminescence spectra from InGaN multiquantum wells heavily doped with Si," *Appl. Phys. Lett.* **72**, 3329 (1998).
52. D. A. Cohen, T. Margalith, A. C. Abare, *et al.*, "Catastrophic optical damage in GaInN multiple quantum wells," *Appl. Phys. Lett.* **72**, 3267 (1998).
53. S. Nakamura, M. Senoh, S. Nagahama, *et al.*, "Continuous-wave operation of InGaN/GaN/AlGaIn-based laser diodes grown on GaN substrates," *Appl. Phys. Lett.* **72**, 2014 (1998).

Properties of III-Nitrides Grown on Si(111) Substrates by Plasma-Assisted Molecular Beam Epitaxy

M. A. Sánchez-García, E. Calleja, F. B. Naranjo, F. Calle, F. J. Sánchez,
and E. Muñoz

Dpto. Ingeniería Electrónica, ETSI Telecomunicación, Universidad Politécnica de Madrid, Ciudad Universitaria s/n, 28040 Madrid, Spain

1. Introduction

III-Nitrides (GaN, AlN and InN) have proven to be excellent materials in visible to ultraviolet (UV) optoelectronic and high temperature electronic applications.¹ The wurtzite nitride polytypes form a continuous alloy system whose direct room temperature bandgaps range from 6.2 eV in AlN to 3.4 eV in GaN to 1.9 eV in InN. This large bandgap range, together with high thermal and chemical stability, material hardness, and thermal conductivity makes nitrides promising for high temperature/power electronics and optoelectronics. The visible to ultraviolet (UV) emission capability of this material system has direct applications in light emitters, as well as integrated full-color displays. The same physical properties also make them good candidates for UV detectors, which have a wide range of commercial and military applications, particularly in those areas where the UV component of light needs to be analyzed in the presence of a strong visible or infrared (IR) background.

The lack of a suitable substrate (that is both lattice and thermally matched) on which to grow these materials has been the main barrier the crystal growers have faced in the last 10 years to produce device-quality material. Though researchers are actively pursuing bulk nitride growth, much work remains to be done before GaN substrates are commercially available. Most of the current achievements in the III-nitrides devices field have been obtained using sapphire and SiC substrates.¹⁻⁵ However, there is still well-founded interest to grow these materials on Si. One first reason would be the integration of the optoelectronic capabilities of the nitrides with the established Si processing and technology. Other related advantages include low expense, wide range of available electrical conductivities, cleavability (easy laser technology), and the possibility for *in-situ* homoepitaxy to start with an atomically smooth and clean Si substrate surface.

In this work we present the properties of GaN, AlN and AlGaIn layers grown on Si(111) substrates by plasma-assisted molecular beam epitaxy (MBE). The growth optimization process will be described before analyzing the morphological, optical, and electrical properties of the different layers. Both undoped and doped material will be studied. It will be shown that the quality of this material is comparable to that grown on sapphire or SiC by MBE.

2. Experimental details

GaN, AlN and AlGaIn layers were grown by MBE using a radio frequency (rf) plasma source to activate the nitrogen and conventional Knudsen effusion cells for the rest of the elements. Details of the growth system and substrate preparation can be found elsewhere.⁶ The quality of the material was assessed using different characterization techniques. The surface morphology was studied by atomic force microscopy (AFM). High resolution X-ray diffraction (XRD) was employed to analyze the structural properties and to determine the ternary composition. Low temperature photoluminescence (PL) experiments were carried out in a double-cycle He-flow cryostat, at temperatures down to 3.6K, using 334 nm line Ar⁺ laser excitation. Scanning electron microscopy (SEM) was performed to observe the morphological structure of the layers and relate them to other properties (e.g., optical and electrical).

3. GaN and AlGaIn growth optimization

The high lattice (17%) and thermal expansion coefficient mismatches present between Si(111) and GaN makes the epitaxial growth a difficult task. A general approach common to other substrates consists in the use of an AlN buffer layer to improve the quality of the GaN films.^{7,8} Therefore, the optimization of the growth of AlN had to be performed before optimizing the GaN.

For the growth of AlN, an initial Al coverage of the Si substrate is needed to avoid the formation of amorphous Si_xN_y at the interface. Using this approach, abrupt and clear interfaces with no sign of amorphous material were grown. High quality AlN layers were obtained using a III/V ratio close to stoichiometry and a high growth temperature (> 850 °C).⁹ Growth under slightly Al-rich conditions provides XRD values (FWHM) of 10 arcmin and surface roughness of 4.8 nm.⁹

GaN films were grown at 750 °C using these optimized AlN buffer layers. The quality of the GaN layer was highly dependent on the quality of the buffer layer. A typical growth rate of 0.52 μm/hr was achieved. The effect of the growth rate at the beginning of the growth was investigated by inspecting the surface roughness by AFM. Figure 1 shows AFM photographs of two GaN layers just differing in the growth rates at the beginning of the growth. In Fig. 1, the sample on the left was grown at a constant rate (0.52 μm/hr) for two hours, while the sample on the right started at a slow growth rate (0.08 μm/hr) and then continued at the normal rate (0.52 μm/hr). Although the crystal quality of both samples was similar in terms of XRD data (10 arcmin), Fig. 1 indicates that the slow-start sample presents a larger coalescence of the grains at the surface. This strategy promotes a higher nucleation and leads to more compact surfaces at the end of the growth. Following this kind of procedure, we have obtained our best FWHM result of 8.5 arcmin and surface roughness of 5.7 nm for a GaN sample.¹⁰

The morphology of the GaN layers was also observed by SEM. It was found

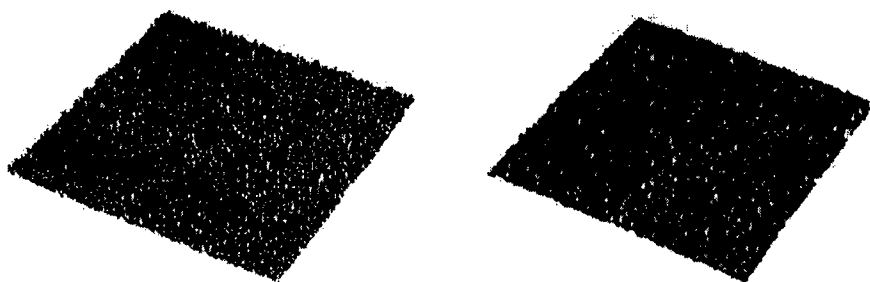


Figure 1. AFM photographs of GaN surfaces: grown at 0.52 $\mu\text{m/hr}$ for two hours (left) and grown initially at a slow 0.08 $\mu\text{m/hr}$ rate, followed by normal growth. (right). Scan areas are 20 X 20 μm .

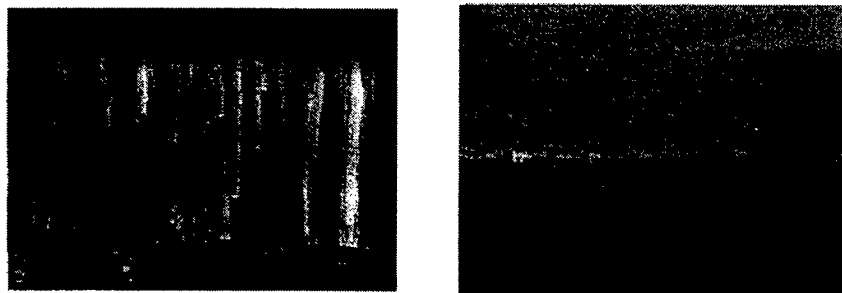


Figure 2. Cross-sectional SEM micrographs of GaN layers grown under different conditions: N-rich (left) and slightly above stoichiometry (right).

that the morphology is highly dependent on the III/V ratio used during growth. Figure 2 shows cross sectional SEM photographs of samples grown with different III/V ratios. In Fig. 2, the left photograph corresponds to a sample grown under N-rich conditions. The layer consists of an array of whisker-like micro-crystals of approximately 60 nm width completely isolated from one another. This columnar morphology evolves to compact films as the III/V ratio is increased to slightly above stoichiometry (Ga-rich condition), as shown in Fig. 2 on the right. As an intermediate case some samples were grown starting with a III/V ratio slightly above 1 (i.e. Ga-rich) and after two hours of growth the Ga flux was reduced to continue growing under N-rich conditions. The result was the appearance of microcrystals on top of the initial layer. This sharp transition indicates the critical role of the III/V ratio used during growth. Such behavior was also observed in GaN layers grown directly on Si with no intermediate AlN buffer layers.

The different morphologies observed in Figure 2 also lead to different optical properties. The low temperature PL spectra of Fig. 3(a) corresponds to the highly columnar sample of Fig. 2 (left). Emission observed from this sample was very intense and excitonic, characteristic of a high-quality material, indicating that each one of those microcrystals is a dislocation-free relaxed crystal.¹¹ The compact film of Fig. 3(b) presented a lower intensity spectra with a single dominant transition characteristic of a material under biaxial residual strain. This decrease in PL intensity when the layer becomes compact can be explained by the generation of a large number of dislocations that behave as nonradiative recombination centers.

Finally, AlGaIn layers were grown with Al content ranging from 10% to 76% showing smooth surfaces for low Al content (up to 20%). Above this Al content the surface morphology becomes rough, indicating that more growth parameter optimization is needed for such layers. The Al content up to 20% was determined by XRD data and compared with low-temperature PL measurements shown in Fig. 4 (films with higher Al content could not be excited optically with our $\lambda = 334$ nm laser source). The results agree well with previous data in the literature.¹² The use of these films as buffer layers for the growth of GaN is currently underway and is expected to further improve the quality of the GaN material.

To summarize this section, optimal GaN layers (8.5 arcmin XRD-FWHM) were obtained using AlN buffer layers grown at 850 °C. The growth of the GaN layer was performed at 750 °C with III/V ratio slightly above stoichiometry showing compact morphology and smooth surfaces (5.7 nm roughness). The III/V ratio during growth is the governing parameter of the morphology of the film. AlGaIn layers with smooth surfaces and intense PL were also grown for Al content below 20%, while more growth optimization is needed for layers with higher Al content.

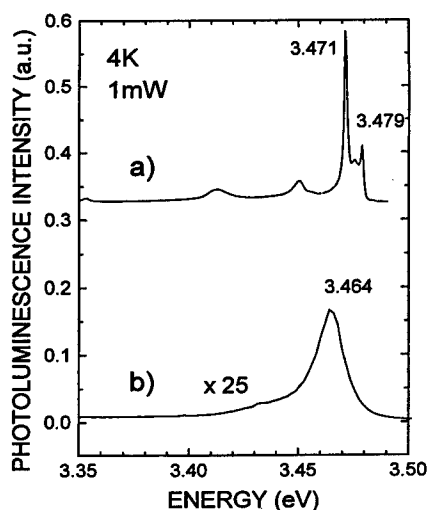


Figure 3. Low-temperature PL spectra of GaN layers grown under different III/V ratios corresponding to Fig. 2.

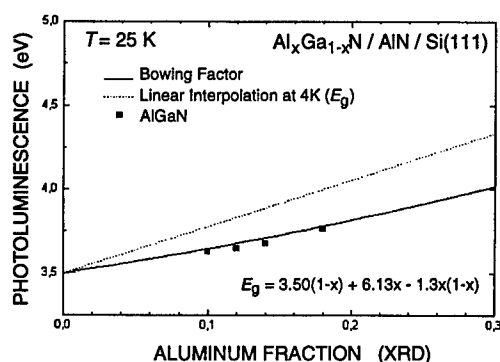


Figure 4. Low-temperature photoluminescence emission of AlGaIn layers with Al content determined by XRD data. The fitting formula for E_g is from Ref. 11.

4. Transport properties of undoped GaN

All unintentionally doped GaN layers reported to date have had n -type background carrier concentrations reaching values up to 10^{18} cm^{-3} in some cases. The origin of this residual concentration is not yet completely understood although native defects (i.e. nitrogen vacancies) and/or oxygen contamination have been proposed as the most probable candidates, depending on the defect density.¹³⁻¹⁵ The reduction of this residual conductivity below 10^{17} cm^{-3} is needed to obtain efficient p -type doping with low compensation and high mobilities. With improved crystal growth techniques, several groups have succeeded in reducing the background electron concentration to 10^{16} cm^{-3} .¹⁶

Transport properties are normally determined from Hall data. However, the reliability of this technique depends on many factors. The layer morphology (columnar, grain boundaries, mixed cubic/hexagonal phases) might restrict parallel conduction. The formation of p -type inversion layers at the interface can also mask the actual conductivity of the GaN layer (Ga and Al behave as acceptors in Si). This phenomenon has already been reported in AlN/Si (111) interfaces.¹⁷

Hall measurements were first performed at room temperature in undoped samples *without* AlN buffer layers. Figure 5 shows the carrier concentration and conductivity type as a function of the Ga flux and the growth temperature. All samples grown at 660°C are n -type with increasing electron concentrations, from 10^{16} to 10^{20} cm^{-3} , as a function of the Ga flux. We found that a III/V ratio reduction leads to a sharp conductivity decrease and to semi-insulating layers. But this effect might partially arise from restricted parallel conduction due to the columnar morphology arising from N-rich growth (see SEM photograph in Fig. 2 on the left). When the growth temperature increases to 720°C the conductivity changes to p -type, independently of the Ga flux as shown in Fig. 5. For a given growth temperature, the longer the growth time, the higher the p -type conductivity; and a further increase in the growth temperature to 770°C leads to

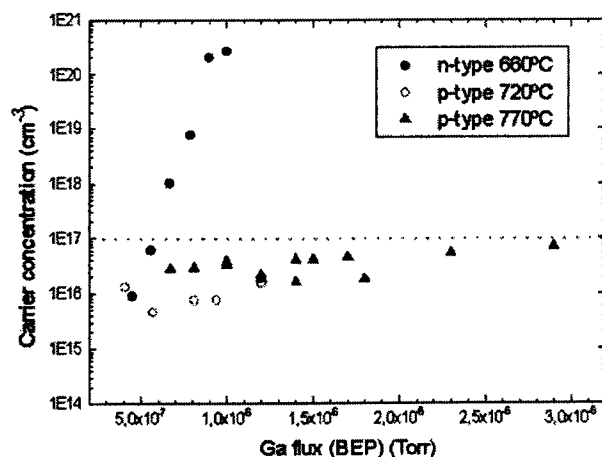


Figure 5. Room-temperature Hall carrier concentration vs. Ga flux for undoped GaN samples grown at different temperatures without AlN buffers. Above 660 °C all samples are *p*-type.

even higher conductivity values, always *p*-type.

The sharp conductivity increase with the Ga flux shown in Fig. 5 for the samples grown at around 660 °C, together with the fact that no conceivable contaminant source in the MBE system would reach such a high level (RGA did not show any trace of O₂ or other possible contaminants), indicate that point defects like V_{Ga} or Ga_i, or complexes involving them, are most likely to be the origin of the *n*-type conductivity in our undoped GaN samples.

Undoped GaN samples grown by MBE above 660 °C show a *p*-type conductivity, with an activation energy E_A of 62 ± 3 meV independent of the III/V ratio, growth temperature and layer thickness, pointing to a common acceptor. This *p*-type conductivity also increases when the ohmic contacts are annealed at higher temperatures or when the layers are thinned by dry etching with SF₆,¹⁸ revealing that the closer to the GaN/Si(111) interface, the higher the *p*-type conductivity. These observations plus the fact that Ga is a shallow acceptor in Si, at around 65 meV above the valence band, leads to the conclusion that the Hall conductivity is dominated by a highly *p*-type interface channel generated by Ga diffusion into the Si substrate. This interpretation is also confirmed by secondary ion mass spectroscopy (SIMS) as shown in Fig. 6. The Ga and Si profiles are shown near the interface. A clear Ga diffusion into the Si substrate, as well as of Si into the GaN layer, is observed.¹⁹

As we have already mentioned, the quality of the GaN layer improves considerably when grown on optimized AlN buffer layers. Because Al behaves also as an acceptor in Si and its diffusion coefficient is even higher than that of Ga, a diffused *p*-type layer at the AlN/Si(111) interface is also expected. Hall data

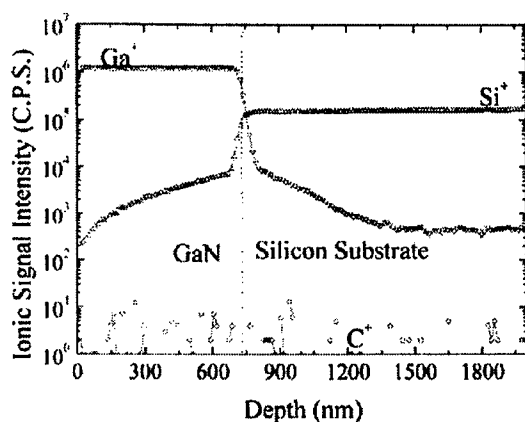


Figure 6. SIMS profile of the GaN/Si interface for a sample grown at 780 °C.

reveal that GaN layers grown on AlN buffer layers, as well as the AlN buffers alone, have *p*-type conductivities with an activation energy very similar to that of the Al acceptors (57 meV) in Si.¹⁹

We employed a new approach to determine the residual carrier concentration of our undoped material grown on AlN buffer layers at growth temperatures above 660 °C. *C-V* measurements yielded values of $N_D = 1.8 \times 10^{17} \text{ cm}^{-3}$ using horizontal Schottky barriers made with Au and ohmic contact with Ti/Al. The same results were obtained when the Si substrate was removed by wet etching with $\text{H}_3\text{NO}:\text{HF}$, indicating that the diffused layer has no effect on the Schottky behavior.

5. Doping of GaN

The achievement of *p*-type doping has been one of the most investigated subjects in the nitrides field. The high *n*-type residual concentration present in the undoped material makes the *p*-type doping a difficult task to start with. On the other hand, controlled *n*-type doping poses no problem and Si is generally used as the dopant.

GaN samples doped with Si were grown at a substrate temperature of 750 °C reaching carrier concentrations up to $1.7 \times 10^{19} \text{ cm}^{-3}$. For cell temperatures at or above 1050 °C, a clear *n*-type conductivity was measured with carrier concentration ranging from $2 \times 10^{18} \text{ cm}^{-3}$ to $1.7 \times 10^{19} \text{ cm}^{-3}$ and mobilities between 20–100 cm^2/Vs . For cell temperatures below 1000 °C, *p*-type conductivity was measured due to the interdiffusion problem mentioned in the previous section.

Three different *p*-type dopants were used: Be, Mg and C. The inset of Fig. 7 shows SIMS results from several GaN layers doped with Be, indicating that the incorporation of Be increases with the Be cell temperature. Hall measurements of these samples were not reliable due to the interdiffusion problem. However, low temperature PL experiments reveal the presence of at least two emissions related

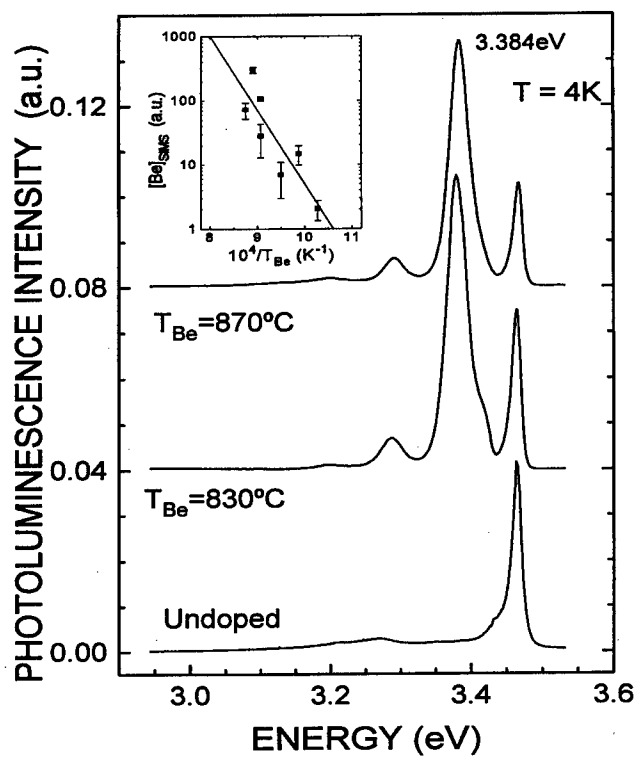


Figure 7. Low-temperature PL of Be-doped GaN layers. Inset shows the Be signal in the SIMS profile of Be-doped GaN layers as a function of Be-cell temperature.

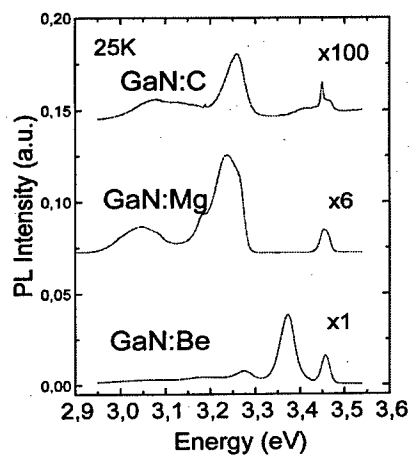


Figure 8. Low-temperature PL of GaN layers doped with Be, Mg, and C.

to Be doping: a wide emission centered around 2.3 eV already reported by other authors,²⁰ and a dominant emission at 3.384 eV (Fig. 7). Temperature and power dependence as well as time-resolved PL experiments of this last emission indicate a donor-acceptor pair character, with an acceptor energy between 90–100 meV.²¹

Doping with Mg and C has been recently started and still needs some optimization and careful study. Hall measurements pose the same reliability problem as before. For comparison with Be-doping, Fig. 8 shows low temperature PL spectra of samples doped with Be, Mg and C. Observing these three spectra, we can conclude that Be has the lowest optical activation value for the acceptor.

6. Conclusions

To summarize, GaN layers have been successfully grown on Si(111) substrates by plasma-assisted MBE, obtaining results comparable to those observed in layers grown on sapphire and SiC by MBE. The morphology of the film is very dependent on the growth conditions employed. Highly columnar morphologies give rise to very intense and excitonic photoluminescence, but poor parallel transport properties, which are needed in devices. An interdiffusion process taking place at the GaN/Si interface prevents us from obtaining reliable Hall data due to the generation of a highly *p*-type conductive channel. Silicon has been used for *n*-type doping, reaching values up to $2 \times 10^{19} \text{ cm}^{-3}$; whereas *p*-type doping with Be, Mg and C has been observed by photoluminescence techniques. Although some final optimization is needed in the Mg and C doping, we can derive a much lower value for the optical activation energy of the Be acceptors (between 90 and 100 meV) compared with Mg and C (around 250 meV).

References

1. S. Nakamura and G. Fasol, *The Blue Laser Diode — GaN Based Light Emitters and Lasers*, Heidelberg: Springer-Verlag, 1997.
2. S. Tanaka, R. S. Kern and R. F. Davis, "Initial stage of aluminium nitride film growth on 6H-silicon carbide by plasma-assisted, gas source molecular beam epitaxy," *Appl. Phys. Lett.* **66**, 37 (1995).
3. B. N. Sverdlov, G. A. Martin, H. Morkoç, and D. J. Smith, "Formation of threading defects in GaN wurtzite films grown on nonisomorphic substrate," *Appl. Phys. Lett.* **67**, 2063 (1995).
4. M. E. Lin, B. Sverdlov, G. L. Zhou and H. Morkoç, "A comparative study of GaN epilayers grown on sapphire and SiC substrates by plasma-assisted molecular-beam epitaxy," *Appl. Phys. Lett.* **62**, 3479 (1993).
5. T. George, W. T. Spike, M. A. Khan, J. N. Kuznia and P. Chang-Chien, "A microstructural comparison of the initial growth of AlN and GaN layers on basal plane sapphire and SiC substrates by low pressure metalorganic chemical vapor deposition," *J. Electron. Mater.* **24**, 241 (1995).

6. M. A. Sánchez-García, E. Calleja, E. Monroy, *et al.*, "The effect of the III/V ratio and substrate temperature on the morphology and properties of GaN and AlN layers grown by molecular beam epitaxy on Si(111)," *J. Cryst. Growth*, **183**, 23 (1998).
7. J. N. Kuznia, M. A. Khan, D. T. Olson, R. Kaplan and J. Freitas, "Influence of buffer layers on the deposition of high quality single crystal GaN over sapphire substrate," *J. Appl. Phys.* **73**, 4700 (1993).
8. S. Yoshida, S. Misawa and S. Gonda, "Improvements on the electrical and luminescent properties of reactive molecular beam epitaxially grown GaN films by using AlN-coated sapphire substrates," *Appl. Phys. Lett.* **42**, 427 (1983).
9. E. Calleja, M. A. Sánchez-García, E. Monroy, *et al.*, "Growth kinetics and morphology of high quality AlN grown on Si(111) by plasma-assisted molecular beam epitaxy," *J. Appl. Phys.* **82**, 4681 (1997).
10. M. A. Sánchez-García, E. Calleja, F. J. Sánchez *et al.*, "Growth optimization and doping with Si and Be of high quality GaN on Si(111) by molecular beam epitaxy," *J. Electron. Mater.* **27**, 276 (1998).
11. F. Calle, F. J. Sánchez, J. M. G. Tijero, M. A. Sánchez-García, E. Calleja, and R. Beresford, "Exciton and donor-acceptor recombination in undoped GaN on Si(111)," *Semicond. Sci. Technol.* **12**, 1396 (1997).
12. H. Angerer, D. Brunner, F. Frendenberg, *et al.*, "Determination of the Al mole fraction and band gap bowing of epitaxial AlGaIn films," *Appl. Phys. Lett.* **71**, 1504 (1997).
13. B. C. Chung and M. Gershenson, "The influence of oxygen on the electrical and optical properties of GaN crystals grown by metalorganic vapor phase epitaxy," *J. Appl. Phys.* **72**, 651 (1992).
14. J. Neugebauer and C. G. Van de Walle, "Atomic geometry and electronic structure of native defects in GaN," *Phys. Rev. B* **50**, 8067 (1994).
15. P. Boguslowski, E. L. Briggs and J. Bernholc, "Native defects in gallium nitride," *Phys. Rev. B* **51**, 17255 (1995).
16. S. Nakamura, Y. Harada, and M. Senoh, "Novel metalorganic chemical vapor-deposition system for GaN growth," *Appl. Phys. Lett.* **58**, 2021 (1991).
17. X. Zhang, D. Walker, A. Saxler, P. Kung, J. Xu and M. Razeghi, "Observation of inversion layers at AlN-Si interfaces fabricated by metal organic chemical vapour deposition," *Electron. Lett.* **32**, 1622 (1996).
18. D. Basak, M. Verdu, M.T. Montojo, *et al.*, "Reactive ion etching of GaN layers using SF₆," *Semicond. Sci. Technol.* **12**, 1654 (1997).
19. E. Calleja, M. A. Sánchez-García, D. Basak *et al.*, "Effect of Ga/Si inter-diffusion on optical and transport properties of GaN layers grown on Si(111) by molecular beam epitaxy," *Phys. Rev. B* **58**, 1550 (1998).
20. J. I. Pankove and J. A. Hutchby, "Photoluminescence of ion-implanted GaN," *J. Appl. Phys.* **47**, 5387 (1976).
21. F. J. Sánchez, F. Calle, M. A. Sánchez-García, *et al.*, "Experimental evidence for Be shallow acceptor in GaN grown on Si (111) by molecular beam epitaxy," to appear in *Semicond. Sci. Technol.* (1998).

Multi-Wavelength Optical Code Division Multiplexing

C. F. Lam and E. Yablonovitch

*Electrical Engineering Department, UCLA, 405 Hilgard Ave., Los Angeles,
CA 90095-1594, USA*

1. Introduction

In the past 10 or 15 years there has been tremendous interest in applying spread spectrum and code division multiple access (CDMA) concepts to optical communications. CDMA has been successfully deployed in military systems for secure communications and cellular phone systems to make efficient use of the radio spectrum. Among the advantages of CDMA systems are their anti-jamming capability, low probability of detection, and inherent security advantage due to spread spectrum encoding. However, for more general communications purposes, most optical CDMA schemes were not competitive, in terms of system throughput and capacity, with more conventional schemes like wavelength division multiplexing (WDM). Here we describe a multi-wavelength spectrally encoded optical CDMA system that begins to approach WDM performance.

In a CDMA system,^{1,2} the signal is spread over a much wider channel than required for data transmission. Each channel consists of a spread spectrum code signature and is broadcast to all receivers on the same network. The code signature is removed at the receiver by auto-correlation with a matched code. Different channels are encoded with signatures orthogonal to one another and unmatched code signatures are rejected. The receiver needs the correct code in order to recover the signal, which leads to enhanced security.

In an optical CDMA system, encoding and decoding are performed in the optical domain. In fact, optical CDMA was first conceived as a multiple access protocol in an optical local area network (LAN) environment, by making use of the tremendous bandwidth available in the optical fiber, thereby avoiding the electronic processing bottleneck.^{3,4}

Different approaches have been proposed for optical CDMA systems. Electrical field detection is widely used in radio systems while intensity detection is more popular in optical systems due to the difficulties in phase locking and polarization maintenance of optical signals. Since electrical fields are bipolar in nature, spreading code signatures with both positive and negative components can lead to true orthogonality,^{1,2} which is vital for avoiding cross-talk. Positive-only intensity detection schemes, used in most optical communication systems, make code orthogonality and cross-channel cancellation more difficult to achieve.

One example of optical CDMA uses an optical delay-line network to transform an ultra-short optical pulse into a train of low intensity pulses occupying

a bit period.³⁻⁶ A conjugate delay-line network is used at the receiver to reconstruct the original short pulse. Orthogonality is impossible in such systems. In order to reduce the crosstalk, codes with long length and small weight are used, which leads to inefficient use of the available spectrum. Code families used in such systems include optical orthogonal codes,⁷ prime sequence codes,⁸ *etc.* The non-zero cross-correlation of these codes severely limits the bit error rate (BER).

Other approaches to optical CDMA involve frequency domain processing by spectral encoding. A dispersive element is used to decompose the frequency components in a broadband optical signal. The coherent approach^{9,10} uses pseudo-random phase encoding of the spectral components to disperse a sharp narrow spike in the time domain into a broad noise-like low intensity signal. The sharp spike is reconstructed by conjugate phase encoding of the received signal. A non-linear threshold device is used to differentiate between a reconstructed high intensity narrow spike and unmatched low-intensity background interference signal. This approach is very sensitive to phase shifts in the fiber channel. The non-linear detection scheme is also clumsy and cumbersome.

A non-coherent spectral intensity encoding approach has also been proposed and demonstrated.^{11,12} In this approach, the transmitter selectively transmits certain wavelength components for each channel. The receiver uses a pair of complementary spectral filters and balanced detectors to cancel unmatched spectra and transmit the desired spectrum. Bipolar signaling and full orthogonality can be achieved in spite of the absence of a coherent local oscillator. Unfortunately, the BER of this system and most other optical CDMA systems is limited by the speckle noise generated due to the interference of optical waves of the same wavelength coming from different transmitters.

In this paper, we describe a novel optical CDMA approach that uses phase encoding of a mode-locked laser to create pseudo-random noise patterns. These patterns can be regarded as a carrier transmitted along with modulation side-bands. At the receiver, this encoded carrier is mixed with the encoded modulation side-bands in a double balanced mixer to recover the information. The ultimate performance of this system is shot noise limited, and is analyzed in this paper.

2. Transmitter

A multi-wavelength optical CDMA network architecture is shown in Fig. 1. Each user broadcasts its optically encoded signal to all other users in the network through a star coupler. The mode-locked laser pulses of each user synchronize with the rest of the network through a time synchronization signal generated by the receiver which monitors the star coupler. The pulses should arrive at the star coupler simultaneously from all transmitters. Absolute optical phase synchronization is not needed.

Figure 2 shows the block diagram of the transmitter which uses a mode-locked laser as the optical source.¹³ The output of the mode-locked laser source

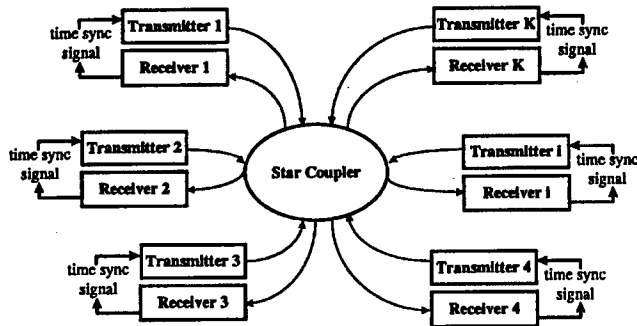


Figure 1. Schematic CDMA network architecture.

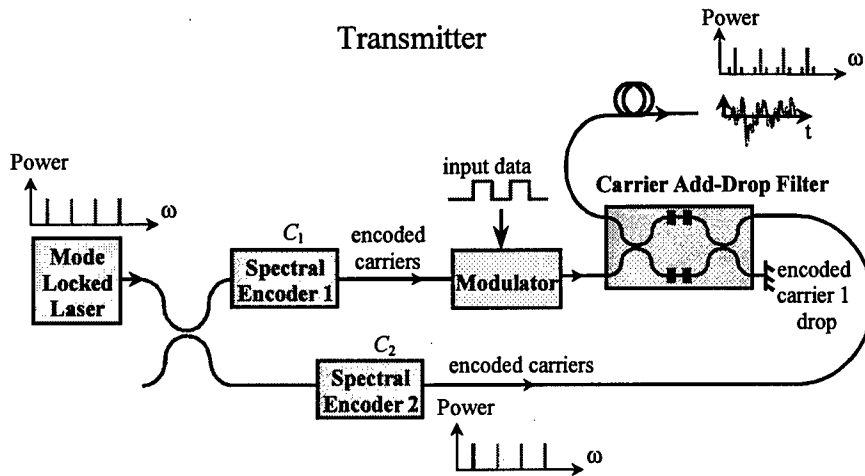


Figure 2. The transmitter provides both encoded side-bands C_1 and a separately encoded carrier C_2 .

can be represented by:

$$E(t) = E_0 \exp(i\omega t) \sum_{n=0}^{N-1} \frac{1}{\sqrt{N}} \exp[in(\Delta\omega)t] \quad (1)$$

where $\Delta\omega/2\pi$ is the pulse repetition frequency.

There are several good reasons to use a mode-locked laser. First, the multiple harmonics of a mode-locked laser are phase-synchronized with respect to each other. They provide the vector space for encoding and are broadband as needed for spread-spectrum CDMA systems. Second, because the phases of the spectral components are locked with respect to each other, the output from a mode-locked laser consists of pulse trains separated by $T = 2\pi/\Delta\omega$. These well defined pulses can be used to synchronize each user with respect to other users. Third, because of the well-defined phase relationships between different spectral components, we

can apply phase encoding to ensure perfectly orthogonal codes, as opposed to other optical CDMA systems which use intensity encoding in the time domain. At the same time, we do not have to track the absolute optical phase, but only the relative phase shift between different mode-locked frequency components.

The mode-locked laser output forms the carrier for the signal. It is split into two halves using a 3-dB splitter. We define the encoding operation $C_k[E(t)]$ for encoder k as:

$$C_k[E(t)] = C_k(t) = E_0 \exp(i\omega t) \sum_{n=0}^{N-1} \frac{1}{\sqrt{N}} \exp\{in(\Delta\omega)t + \Phi_{kn}\} \quad (2)$$

where Φ_{kn} is the encoded phase on the n th spectral component. The codes are designed to be orthogonal so that two codes C_k and C_h satisfy:

$$C_k \cdot C_h^* = E_0^2 \sum_{n=0}^{N-1} \frac{1}{N} \exp\{i(\Phi_{kn} - \Phi_{hn})\} = E_0^2 \delta_{kh} \quad (3)$$

where $\delta_{kh} = 1$ if $k = h$ and 0 otherwise. In the simplest case, if we use 0's and π 's as the encoded phase shifts, the codes correspond to multiplication coefficients of +1 and -1 in the frequency domain. All the good bipolar codes developed for radio CDMA can be directly applied to our system including, for example, the Hadamard codes¹⁴ that form the rows of a Hadamard square matrix. Two codes are orthogonal when the phase differences between the corresponding spectral components are given by $\exp[i(\Phi_{kn} - \Phi_{hn})]$ and are uniformly distributed on a unit circle in a complex plane. A new encoder family using cascaded feedback Mach-Zehnder interferometers will be introduced in the latter part of this paper.

For the k th user, the first half of the carrier passes through a spectral filter which phase encodes the carrier components as:

$$C_{k,1}(t) = \frac{1}{\sqrt{2}} E_0 \exp(i\omega t) \sum_{n=0}^{N-1} \frac{1}{\sqrt{N}} \exp\{in(\Delta\omega)t + \Phi_{kn,1}\} \quad (4)$$

The other half of the spectral components of the k th user are encoded by a different code $C_{k,2}$ in a similar way:

$$C_{k,2}(t) = \frac{1}{\sqrt{2}} E_0 \exp(i\omega t) \sum_{n=0}^{N-1} \frac{1}{\sqrt{N}} \exp\{in(\Delta\omega)t + \Phi_{kn,2}\} \quad (5)$$

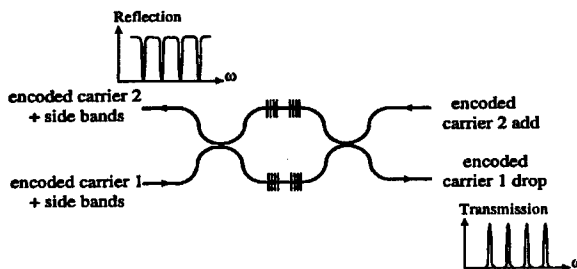


Figure 3.

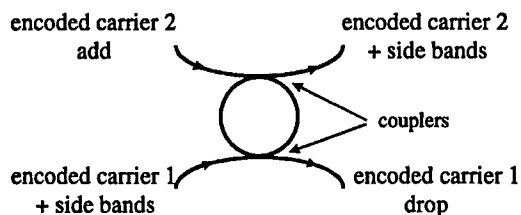


Figure 4.

The power spectral density function of the encoded carrier still looks the same as that of the unencoded carrier in the frequency domain. However, in the time domain, it resembles a pseudo-random noise pattern.

A modulator is placed after the encoded carrier $C_{k,1}$ to impose data modulation. Either phase or intensity modulation could be applied. The modulated output is:

$$X_k(t) = s_k(t)C_{k,1}(t) \quad (6)$$

where $s_k(t)$ is the data signal for the k th user. For amplitude shift keying (ASK) $s_k(t) = 0$ or $+1$, while for phase shift keying (PSK), $s_k(t) = +1$ or -1 . Modulation generates the information-containing side bands $s_k(t)C_{k,1}(t)$ around the pure carrier tones $C_{k,1}(t)$. This signal is passed through a carrier add-drop filter consisting of a balanced Mach-Zehnder interferometer (MZI) with a pair of identical high-finesse Fabry-Perot (FP) filters in both arms. The FP filters have very sharp frequency transmission at the carrier tone frequencies and very high reflection for frequencies other than the carrier tones (Fig. 3). The residual encoded carrier tones $C_{k,1}(t)$ will pass through the MZI and will be dropped at the carrier drop port on the opposite side of the input signal. The information-containing side bands will be reflected to the other port on the same side of the signal.

The transmitter also adds the encoded carrier $C_{k,2}(t)$ to the reflected side bands from the carrier add port of the MZI. The added carrier will be extracted at the receiver as the "local oscillator" to demodulate the transmitted signal. A similar MZI arrangement for carrier add/drop filters has been used with fiber gratings as a wavelength division multiple-access (WDM) add/drop multiplexer.¹⁵ Another possible carrier add/drop filter consists of a ring coupler pair shown in Fig. 4. This filter is made up of two weak couplers connected in a ring whose size matches the resonant condition for the carrier tones. The carriers will be coupled out of the carrier drop port because of resonance. The device can be made very compact and has the potential for integrated photonics.¹⁶

3. Receiver

The side bands and the encoded carrier are broadcast to the receivers through a star coupler. Figure 5 shows the receiver structure. The same carrier add-drop filter is

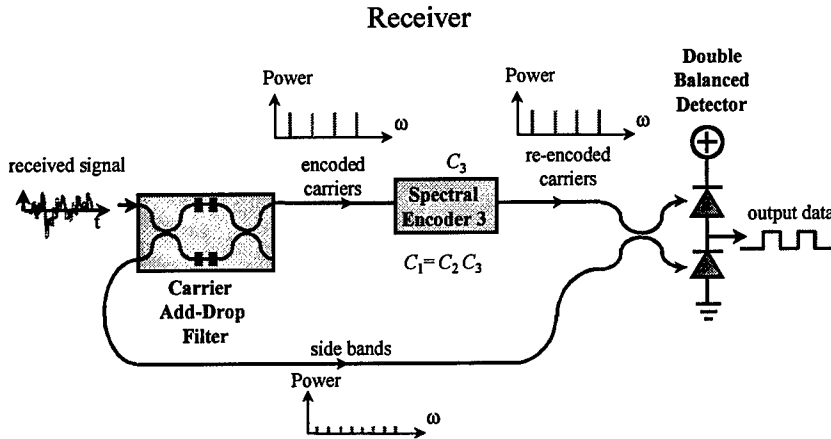


Figure 5.

used to separate the side bands and the encoded carrier tones. The carrier tones go through another spectral encoder C_3 at the receiver such that $C_{k,1} = C_{k,3}C_{k,2}$ and $C_{k,1}$ is orthogonal to $C_{k,3}C_{j,2}$ for $j \neq k$. Thus the receiver reconstructs the encoded carrier $C_{k,1}$. The side bands and the reconstructed carriers beat at a second 3-dB splitter which is part of a double balanced detector.

Without loss of generality, let $k = 1$ be the desired channel. The input to the second beam splitter consists of the multiplexed side bands and the re-encoded carriers. The outputs consist of the sum $R_{U1}(t)$ and difference $R_{L1}(t)$ of the side bands and the encoded carrier of the desired user, i.e. user 1:

$$R_{U1}(t) = \frac{1}{\sqrt{2}} \sum_{k=1}^K s_k(t) C_{k,1}(t) + \frac{1}{\sqrt{2}} C_{1,3} \sum_{k=1}^K C_{k,2}(t) \quad (7)$$

$$R_{L1}(t) = \frac{1}{\sqrt{2}} \sum_{k=1}^K s_k(t) C_{k,1}(t) - \frac{1}{\sqrt{2}} C_{1,3} \sum_{k=1}^K C_{k,2}(t) \quad (8)$$

The photodetectors used in the balanced receiver can be considered a pair of square law devices followed by a low pass filter. Therefore, the output from the balanced detector will be:

$$\begin{aligned} Z_1(t) &= \text{LP} \left[\frac{\Re}{2} |R_{U1}(t)|^2 - \frac{\Re}{2} |R_{L1}(t)|^2 \right] \\ &= \text{LP} \left[\frac{\Re}{2} \left(\sum_{k=1}^K s_k(t) C_{k,1}(t) \right) \cdot \left(C_{1,3} \sum_{k=1}^K C_{k,2}(t) \right)^* \right] \end{aligned} \quad (9)$$

where \Re is the responsivity of the photodetectors. The square terms of the side bands and re-encoded carrier tones are cancelled at the balanced detector output, eliminating any common mode fluctuations. The remaining term is the low-pass filtered cross product between the side bands and the re-encoded carriers. By the

code orthogonality requirement, all the dot product terms are cancelled except for user 1. The operation of $C_{1,3}$ on $C_{1,2}$ regenerates $C_{1,1}$ — the correct "local oscillator" used to extract the user 1 signal $s_1(t)$. By maintaining mode-locked time synchronization, perfect orthogonality can be achieved.

4. Decoder

In this section, we describe the design of an encoder that uses series-connected MZIs. The MZIs are set in a feedback configuration as shown in Fig. 6. The output of the feedback MZI (FBMZI) is determined by three parameters, the feed forward path length difference l_f between the two feed forward paths, the extra phase θ introduced by a phase shifter at the second arm of the MZI and the feedback delay introduced by the feedback path l_b . A general frequency transfer function for an arbitrary FBMZI can be easily derived using feedback network theory. Here, we describe a specific design that will generate orthogonal codes.

We assume N , the total number of frequency components in the mode-locked laser spectrum, to be a power of 2, i.e. $N = 2^L$, where L is an integer. The feed forward path length difference l_f and the feedback path length l_b can be expressed by the corresponding time delays $\tau_f = l_f/c$ and $\tau_b = l_b/c$, where c is the speed of light in the waveguide. For simplicity, we restrict the values of τ_f and τ_b to integral multiples of a basic time unit $\tau_0 = 2\pi/(N\Delta\omega)$, which is the reciprocal of the total CDMA bandwidth. In addition, let us restrict the phase shift θ to either 0 or π .

It can be easily shown¹⁷ that setting the feed forward path difference delay to be $\tau_f = N\tau_0/2 = \pi/\Delta\omega$ corresponds to an MZI whose free spectral range (FSR) is $2\Delta\omega$. We can view the MZI as a 2×2 switch for the frequency components that

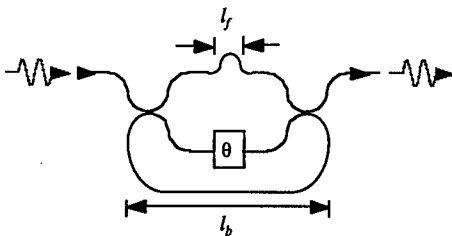


Figure 6.

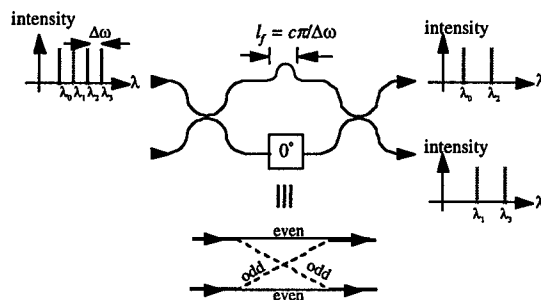


Figure 7.

are separated by $\Delta\omega$. Half of the frequency components (say the even ones) will see the MZI in the through state and the other half (the odd components) will see the MZI in the cross state (Fig. 7). When a phase shift $\theta = \pi$ is introduced at the MZI, the roles of the frequency components are changed so that the ones previously in the through state are now in the cross state and vice versa.

Components in the cross state are fed back to the other input port of the MZI and come out of the same output port as those in the through state after a feedback delay τ_b , which is translated into a phase shift. The 0 and π phase shifts in the feed forward path represent two binary states of an encoder.

Let the feedback delay be $\tau_b = q\tau_0$, where q is an integer. Because of limited space, we will state the following theorems without proof.

- **Theorem 1.** If q and N are co-prime (i.e. they have no common factor), the codes generated by the 0 and π states are orthogonal. The net phase differences for all the N frequency components are distributed on the complex unit circle without repetition.

The actual phase difference pattern is dependent on the value of q , which can be considered a cryptographic factor¹⁸ for encoding. Figure 8 shows the encoded phase differences on the unit circle for $N = 32$ and $q = 1, 3$, and 7 .

- **Theorem 2.** If q and N are not co-prime and have as their greatest common divisor (gcd) the integer 2^m , $m < L$, the codes produced by the 0 and π states are orthogonal within a free spectral range (FSR) of $(N\Delta\omega)/2^m$.

In other words, the encoded phase difference pattern will repeat for every $N/2^m$ frequency components and the orthogonality also covers every $N/2^m$ frequency components. For simplicity, we will say the normalized FSR of the encoder is $N/2^m$ later on. Of course, the orthogonality extends to all the N components. We can express q as $q = 2^m r$ where r is an odd number that determines the actual locations of the frequency components on the complex plane. Figure 9 shows the distribution of the phase differences for $r = 3$ and 7 , $N = 32$, and $m = 2$.

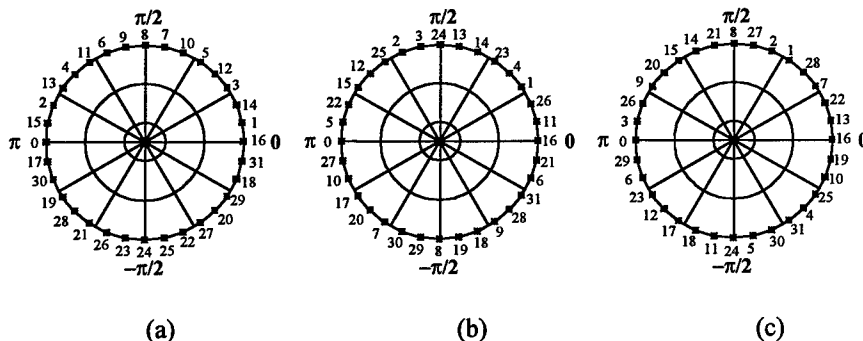


Figure 8. Distribution of the phase difference between two different states for an encoder with (a) $q = 1$, (b) $q = 3$ and (c) $q = 5$. $N = 32$ frequencies labeled from 0 to 31 are used for encoding.

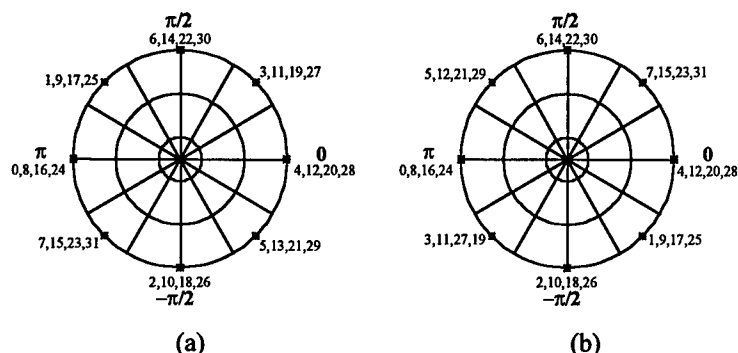


Figure 9. Distribution of the phase difference between two different states for an encoder with $q = 2^m r$ where $m = 2$ and (a) $r = 3$ and (b) $r = 7$. $N = 32$ frequencies labeled from 0 to 31 are used for encoding.

Theorems 1 and 2 together state that for all the integers q , the two states formed by the FBMZI encoder are orthogonal to each other. The next theorem states that a family of orthogonal codes can be generated by cascading encoders with different FSRs.

- **Theorem 3.** Suppose two FBMZIs have a normalized FSR of N and $N/2$ respectively. Each stage has two states 0 and π . The four possible states obtained by combining the two stages generate four mutually orthogonal codes.

Theorem 3 can be extended by induction to prove that by cascading a series of encoders with L different normalized FSRs (i.e. $2, 2^2, \dots, 2^L$) one can obtain all 2^L orthogonal codes. The 2^L codes are produced by different combinations of the 0 or π shifts in the various stages. One plausible way of cascading the FBMZIs is to use a waveguide design with "concentric" circles touching each other to form the 3-dB couplers. The waveguides are nested in cascade as shown in Fig. 10. The output waveguide cut the loops perpendicularly to eliminate the crosstalk.

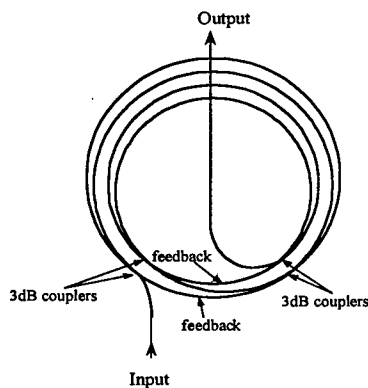


Figure 10.

5. Performance analysis

In this section, we investigate the performance limit of the proposed optical CDMA system. Again, we take user 1 as the intended channel. Assume the carriers $C_{k,1}$ and $C_{k,2}$ have the same amplitude. The value of $s_k(t)$ is 0 or 1 for ASK and -1 or $+1$ for PSK. Ideally, the carrier decoding by the $C_{k,3}$ does not change the amplitude of $C_{k,2}(t)$ because only phase encoding is used.

From Eqs. (4) and (5), we obtain that for each user the power in both carriers $P[C_{k,1}(t)]$ and $P[C_{k,2}(t)]$ is $E_0^2/4$. Assume 0 and 1 bits are equally probable. The average received power is:

$$P_{rec} = \alpha P[C_{k,1}(t)] + P[C_{k,2}(t)] = \frac{(\alpha + 1)E_0^2}{4} \quad (10)$$

or $E_0 = [4P_{rec}/(\alpha + 1)]^{1/2}$, where $\alpha = 1/2$ for ASK and $\alpha = 1$ for PSK. Here we have assumed that the received power from all the users are equal. The splitting loss at the star coupler is also ignored.

The useful output signal is given by Eq. (9). Due to the code orthogonality, only the matched term corresponding to user 1 will survive. So we have:

$$\begin{aligned} Z_1(t) &= \text{LP} \left[\frac{\Re}{2} \left(\sum_{k=1}^K s_k(t) C_{k,1}(t) \right) \bullet \left(C_{1,3} \left[\sum_{k=1}^K C_{k,2}(t) \right] \right)^* \right] \\ &= \text{LP} \left[\frac{\Re}{2} (s_1(t) C_{1,1}(t)) \bullet (C_{1,3} [C_{1,2}(t)])^* \right] \\ &= \text{LP} \left[\frac{\Re}{2} (s_1(t) C_{1,1}(t)) \bullet (C_{1,1}(t))^* \right] \\ &= \frac{\Re E_0^2}{4} s_1(t) = \frac{\Re P_{rec}}{\alpha + 1} s_1(t) \end{aligned} \quad (11)$$

The average received signal (root mean square) is thus:

$$I_{sig} = \sqrt{\overline{Z_1(t)^2}} = \frac{\sqrt{\alpha}}{\alpha + 1} \Re P_{rec} \quad (12)$$

The signals detected by each of the two photodetectors in the balanced receiver are given by $R_{U1}(t)$ and $R_{L1}(t)$ as in Eqs. (7) and (8). Since the received user power (both side bands and re-encoded carriers) is split equally by the 3dB splitter in front of the balanced detector, the average optical power seen by each photodetector is:

$$\frac{1}{2} \overline{|R_{U1}|^2} = \frac{1}{2} \overline{|R_{L1}|^2} = \frac{K P_{rec}}{2} \quad (13)$$

This detected optical power will contribute shot noise¹⁷ at each photodetector. The shot noise produced is therefore:

$$\langle I_{sh}^2 \rangle_{U1} = \langle I_{sh}^2 \rangle_{L1} = 2qIB_d = q\Re K P_{rec} B_d \quad (14)$$

where \mathfrak{R} is the responsivity of the photodetector, B_d is the data bandwidth and I is the average photocurrent produced at each photodetector. The total shot noise is thus:

$$\langle I_{sh}^2 \rangle = \langle I_{sh}^2 \rangle_{U1} + \langle I_{sh}^2 \rangle_{L1} = 2q\mathfrak{R}KP_{rec}B_d \quad (15)$$

We can see that co-channel users degrade the performance of the system by increasing the shot-noise proportionately. For a received power of -20 dB ($10\mu\text{W}$) per active user at a receiver bit rate of 1 Gbps, using $\mathfrak{R} = 0.8\text{A/W}$ for a typical InGaAs PIN photodiode working in the $1.5\mu\text{m}$ wavelength range, the rms shot noise is 5.06×10^{-8} A for one user.

Another source of noise is the Gaussian-distributed thermal noise from the receiver pre-amp which is given by:

$$\langle I_{th}^2 \rangle = \frac{4kT}{R_L} B_d = 8\pi kTB_d^2 C \quad (16)$$

where R_L is the receiver load resistance and $B_d = 1/T_b = 1/2\pi R_L C$ (T_b is the bit period and C is the load capacitance). For state-of-the-art technology, $C \approx 0.02$ pF. Using other parameter values as before, the rms thermal noise is 4.57×10^{-8} A.

It is worth noting that all signals are split equally between the two photodetectors in the balanced detector, so the common-mode fluctuations in the signal and carrier power are cancelled by the balanced detector, except for the shot noise. Since shot noise and thermal noise arise from independent mechanisms, the total mean square noise $\langle I_n^2 \rangle$ is their sum.

The bit error rate (BER) for ASK is given by:¹⁹

$$BER = \frac{1}{2} \left[1 - \text{erf} \left(\sqrt{\frac{\gamma E_b}{2N_0}} \right) \right] \quad (17)$$

where E_b is the average bit energy, $\gamma = 1$ for ASK and $\gamma = 2$ for PSK, and N_0 is the two sided noise power spectral density. One then obtains:

$$\frac{E_b}{N_0} = \frac{I_{sig}^2 T_b}{N_0} = \frac{I_{sig}^2}{N_0 B_d} = \frac{I_{sig}^2}{\langle I_n^2 \rangle} \quad (18)$$

Figure 11 plots the BER against the number of co-channel users at various received power levels for PSK (a 1 Gbps data rate is assumed for each user). It is seen from Fig. 11 that for -20 dBm received power per channel, 197 and 346 concurrent users are allowed for a BER of 10^{-15} and 10^{-9} respectively.

It is also worth noting that this is the worst-case situation. For a system with bursty traffic, the number of concurrent users will likely to be much less than the total number of subscribers. From the above analysis, the SNR is shot-noise limited when the total received power $P_{rec}K$ is large and is approximately:

$$SNR \approx \frac{I_{sig}^2}{\langle I_{sh}^2 \rangle} = \frac{\alpha \mathfrak{R} P_{rec}}{2q(\alpha + 1)^2 K B_d} \quad (19)$$

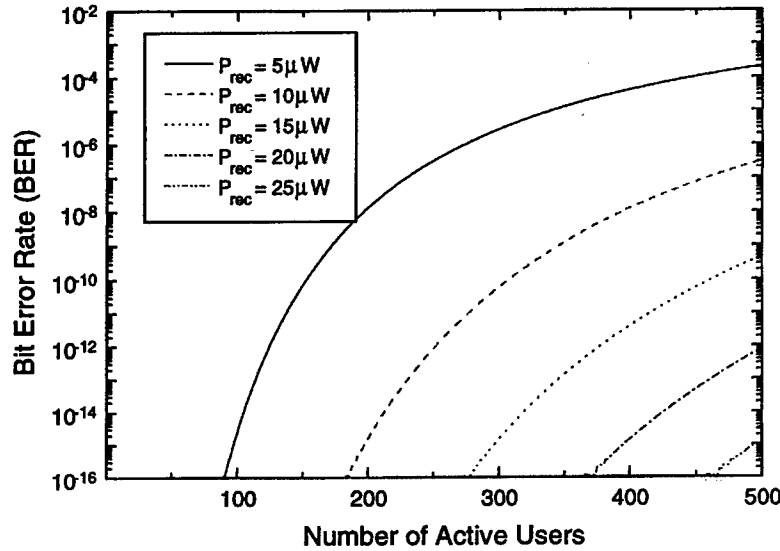


Figure 11. BER vs. number of active users for PSK.

The total network throughput is given by KB_d . The network throughput scales linearly as the received power per channel increases.

From the transmitter's point of view, the signal is broadcast to all the users in the network by a passive optical star coupler. Suppose the network size is the same as the total number of active users (the most pessimistic case). Then the splitting loss is $10\log K$ dB. Neglecting the transmission loss and other non-idealities, the signal to noise ratio in terms of the transmitter power P_t is given by:

$$SNR \approx \frac{I_{sig}^2}{\langle I_{sh}^2 \rangle} = \frac{\alpha \mathcal{R} P_t}{2q(\alpha+1)^2 K^2 B_d} \quad (20)$$

The total throughput is given by:

$$KB_d \approx \frac{\alpha \mathcal{R} P_t}{2q(\alpha+1)K \cdot SNR} \quad (21)$$

For a given available transmitter power, to obtain higher throughput, the total number of users needs to be smaller, which also means that each channel needs to handle a bigger bandwidth. Assuming 10 mW available optical power at the transmitter output and 150 concurrent users, the network can support a total capacity of 2.3 THz for PSK at 10^{-9} BER. Figure 12 plots the achievable throughput for different values of P_t . The transmitter power is assumed to be evenly distributed among all the subscribers who are all active at the same time (again the worst case).

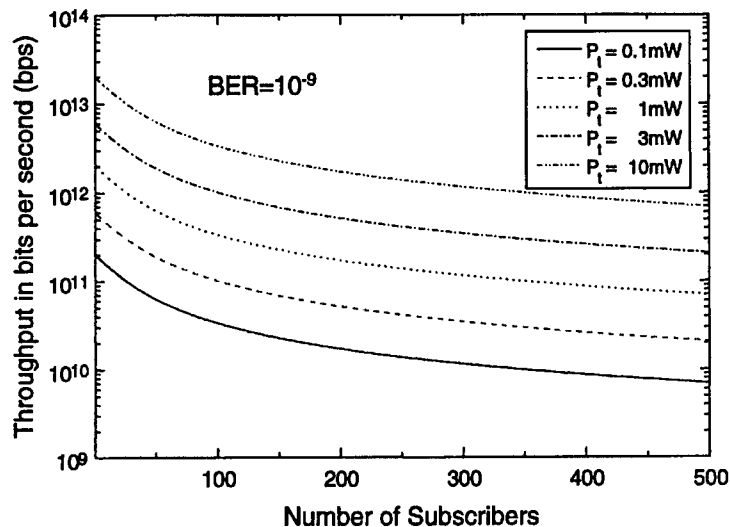


Figure 12. Throughput vs. number of subscribers for PSK.

6. Discussion and conclusion

In the proposed system, we have inserted a separately encoded carrier with the modulated signal for transmission. This transmitted carrier is separated from the information bearing side bands at the receiver and is used to reconstruct the code that is applied to the information-bearing carrier. This separately encoded carrier serves as an externally supplied local oscillator at the receiver, eliminates the requirement of a separate local oscillator, and helps to reduce the system cost. Since the encoded pure carrier and the side bands are generated from the same source and they travel through the same network to the receiver, there is no complicated phase locked loop mechanism required for phase synchronization of the local oscillator. The idea is stimulated by amplitude modulated (AM) radio communication systems. The output spectrum of an AM signal consists of side bands and the carrier tone. The envelope detector in an AM receiver mixes the side bands with the carrier tone, which functions as an externally supplied local oscillator in exactly the same fashion as our balanced photodetectors.

The codes applied to the pure carrier and the sides bands are orthogonal to each other so that simple square law detection will not give any output, protecting the signal from being intercepted by the unwanted receivers and enhancing the system security. As an alternative, one could conceive transmitting the unencoded carrier, which is produced by the mode-locked laser, to reconstruct the code used on the side bands. Since each user broadcasts its signal to all the users in the network, different copies of the unencoded carrier will add non-coherently at the receiver and fluctuate in intensity. This fluctuation is due to the difficulty of synchronizing the absolute phase of the optical carrier. The reconstructed code

will therefore fluctuate in the same way as the laser speckle noise. The beating of this noisy code with the side bands will give excess fluctuations in the output signal. Thus, by encoding the pure carriers in each channel with a different code, only the desired encoded carrier will detect the desired side bands. The carrier from all other channels will carry orthogonal codes.

Time synchronization of mode-locked pulses is required, however, and has been researched in studies of high-speed optical TDMA systems. The synchronization techniques used in those systems could be applicable to our CDMA system. There are two methods to maintain time synchronization among the transmitters. In the first approach, a high performance mode-locked laser source may be shared across the network by distributing its outputs to various users using the star coupler. Since the cost of the high performance source is shared among different users, the average cost could still be reasonable. In the second approach, a master reference laser can be used to synchronize the individual sources through the star coupler.

We have seen that the balanced receiver used in this paper has the ability to reject the common-mode fluctuation in the signal sources and achieve true orthogonality. Eventually, the system will be shot noise limited when the number of users accessing the network is large. When the system is shot noise limited, increasing the signal power improves the BER by increasing the SNR. However, when a system is speckle noise limited as in a non-coherent spectral intensity encoded system,^{11,12} the SNR does not improve by increasing the signal power and the system performance is very limited.

In conclusion, we have proposed an optical CDMA system that is able to achieve full orthogonality and shot noise limited performance. An encoder structure based on feedback Mach-Zehnder interferometers has been proposed. The performance analysis suggests an achievable throughput of 1 Tbit/s.

References

1. R. L. Pickholtz, D. L. Schilling, and L. Milstein, "Theory of spread spectrum communications — a tutorial," *IEEE Trans. Commun.* **30**, 855 (1982).
2. R. L. Peterson, R. E. Ziemer, and D. E. Borth, *Introduction to Spread Spectrum Communications*, Englewood Cliffs, NJ: Prentice Hall, 1995.
3. J. Y. Hui, "Pattern code modulation and optical decoding — a novel code-division multiplexing technique for multifiber networks", *IEEE J. Selected Areas Commun.* **3**, 916 (1985).
4. P. R. Prucnal, M. A. Santoro, and T. Fan, "Spread spectrum fiber-optic local area network using optical processing," *J. Lightwave Technol.* **4**, 547 (1986).
5. P. R. Prucnal, M. A. Santoro, and S. K. Sehgal, "Ultrafast all-optical synchronous multiple access fiber optical networks," *IEEE J. Selected Areas Commun.* **4**, 1484 (1986).

6. J. A. Salehi, "Code division multiple-access techniques in optical fiber networks — part I: fundamental principles," *IEEE Trans. Commun.* **37**, 824 (1989).
7. F. R. K. Chung, J. A. Salehi, and V. K. Wei, "Optical orthogonal codes: design, analysis and applications," *IEEE Trans. Info. Theory* **35**, 595 (1989).
8. L. Tancceviski and I. Andonovic, "Hybrid wavelength hopping/time spreading schemes for use in massive optical networks with increased security," *J. Lightwave Technol.* **14**, 2636 (1996).
9. J. A. Salehi, A. M. Weiner, and J. P. Heritage, "Coherent ultrashort light pulse code-division multiple access communication systems", *J. Lightwave Technol.* **8**, 478 (1990).
10. C. C. Chang, H. P. Sardesai, and A. M. Weiner, "Code-division multiple-access encoding and decoding of femtosecond optical pulses over a 2.5-km fiber link," *IEEE Photonics Technol. Lett.* **10**, 171 (1998).
11. M. Kvehrad and D. Zaccarin, "Optical code-division-multiplexed systems based on spectral encoding of noncoherent sources", *J. Lightwave Technol.* **13**, 534 (1995);
L. Nguyen, T. Dennis, B. Aazhang, and J. F. Young, "Optical spectral amplitude CDMA communication," *J. Lightwave Technol.* **15**, 1647 (1997).
12. C. F. Lam, D. T. K. Tong, M. C. Wu, and E. Yablonovitch, "Experimental demonstration of bipolar optical CDMA system using a balanced transmitter and complementary spectral encoding," *IEEE Photonics Technol. Lett.* **10**, 1504 (1998).
13. Y. K. Chen and M. C. Wu, "Monolithic colliding pulse mode-locked quantum well lasers," *IEEE J. Quantum Electronics* **28**, 2176 (1992).
14. A. K. Jain, *Fundamentals of Digital Image Processing*, Englewood Cliffs, NJ: Prentice Hall, 1986.
15. F. Bilodeau, D. C. Johnson, S. Theriault, *et al.*, "An all-fiber dense-wavelength-division multiplexer/demultiplexer using photoimprinted Bragg gratings," *IEEE Photonics Technol. Lett.* **7**, 388 (1995).
16. B. E. Little, J. S. Foresi, G. Steinmeyer, *et al.*, "Ultra-compact Si-SiO₂ microring resonator optical channel dropping filters," *IEEE Photonics Technol. Lett.* **10**, 549 (1998).
17. P. E. Green, Jr., *Fiber Optic Networks*, Englewood Cliffs, NJ: Prentice-Hall, 1993.
18. N. Koblitz, *A Course in Number Theory and Cryptography*, 2nd ed., New York: Springer-Verlag, 1948.
19. R. E. Ziemer and R. L. Peterson, *Introduction to Digital Communication*, New York: Macmillan, 1992.

Photonic Lattices in Semiconductor Waveguides

Jeff F. Young, P. Paddon, V. Pacradouni, T. Tiedje

*Department of Physics and Astronomy, University of British Columbia,
Vancouver, British Columbia, Canada, V6T 1Z4*

S. Johnson

*Center for Solid State Electronics Research, Arizona State University, Tempe, AZ
85287-6202, U.S.A.*

1. Introduction

One way to avoid the beaten path is to scout ahead of it, forging trails that seem best for future paving. In the context of the information highway, these trails are leading towards single-electron devices, hybrid optical integrated circuits, quantum dots, *etc.* The alternative is to set course perpendicular to the beaten path, searching for new materials, devices and systems that may ultimately offer the ability to do things that smaller and more densely packed CMOS or optoelectronic devices cannot. Obvious examples along this path include systems based on neural networks, quantum computers, or all-optical circuits.

It is impossible to predict which new materials, devices and architectures will prove practical and market-worthy alternatives to those outlined in current business plans. It is quite likely that a distinctive feature of any revolutionary new technology will be that it will not rely solely on the ability to control the flow of charge or photons through essentially serial pipelines. A natural strategy is therefore to explore ways in which it might be possible to control the amplitude and phase of electronic and/or photonic flux in extended (parallel) networks. Stated another way, rather than manipulating populations of the "natural" eigenstates of "bulk" materials, means must be developed for designing "new" eigenstates and controlling them directly.

Considerable progress has already been achieved along these lines in the electronic domain. Epitaxial growth and nanofabrication technologies have been used to realize several novel one-dimensional (1D) and 3D heterostructures and devices wherein the shape, size and scattering properties of electronic eigenstates can be controlled to the same qualitative extent that carrier density and current can be controlled in CMOS devices. There has been relatively little done along these lines in the optical domain despite the compelling suggestions of Yablonovitch¹ and John² in 1987 that periodic dielectric lattices offer a powerful means of controlling the vacuum. These authors independently showed how 3D control of dielectric texture on a length scale Λ can be used to dramatically alter the dispersion and related photon density of states at frequencies $\nu \sim c/n_a\Lambda$ where n_a is

the average refractive index of the textured dielectric and c is the speed of light. This approach is particularly powerful because it does not rely on electronic resonances of the host, and therefore the control over photonic eigenstates can be achieved without necessarily introducing loss. Substantial experimental work at microwave frequencies³⁻⁵ has shown that it is possible to produce artificial optical materials — photonic bandgap materials (PBMs) — within which no electromagnetic excitations in a certain frequency range (the photonic bandgap of that structure) are allowed to propagate in any direction. Furthermore, midgap "defect" states within these periodic dielectric crystals can be localized photons that present the only electromagnetic excitation possible at a certain frequency within the PBM. Judicious placement of coupled defect states can be used as a means of building artificial photonic atoms, molecules or superlattices. Either on their own, or integrated with resonant electronic media, these structures offer an intriguing area of research in the context of alternative material hosts for future information processing technologies.

2. Photonic bandstructure materials

Much has been learned about full 3D and 2D PBMs since their invention in 1987. A host of optical bandstructure calculations have been reported⁶⁻⁹ that, by and large, verify experimental studies of the dispersion and stop bands in 3D and 2D crystals of various symmetries and index contrasts. The index or dielectric contrast of a PBM plays the role of the ionic potential in electronic crystals; the larger the dielectric contrast, the stronger the scattering of photons and the larger the effect on the photon dispersion. In general, a photonic lattice with a particular symmetry may or may not support a full photonic bandgap, but if it does, it will only be possible for certain ranges of index contrast and filling fractions.

There has also been considerable experimental and theoretical work on defect states in PBMs.⁸ It has been clearly established that photons can be localized within just a few crystal lattice constants, making even thin slices of such materials very effective high-Q resonators or filters

The equations that govern the photonic dispersion in PBMs scale as $\omega\Lambda/c$. This scaling implies that everything that has been learned in the convenient microwave region of the spectrum can be applied in the near-infrared region if the pitch of the dielectric lattices can be brought down to the 200–500 nm range. Sajeev John¹⁰⁻¹¹ has theoretically explored several fascinating quantum electronic effects that should be manifest in PBMs when electronic dipole resonances are tuned near the optical bandgap. The work on exciton-polaritons in 1D semiconductor resonators over the past 5 years¹²⁻¹³ foreshadows the rich electron-photon physics that will be accessible when PBMs are realized in III-V semiconductor hosts. The fact that the very same 1D cavities used to study these vacuum Rabi coupling phenomena are crucial to the vertical cavity surface emitting laser (VCSEL) industry speaks to the potential applications that await this more general class of active PBMs.

- *Progress in near-infrared PBMs*

The production of high-quality PBMs with lattice constants less than 500 nm is clearly a technological challenge. Several strategies have been suggested and attempted with limited success to date. Full 3D face centered cubic (fcc) structures can in principle be fabricated the same way as the microwave structures were made by Yablonovitch and co-workers.⁵ Cheng and Sherer¹⁴ have succeeded in using an ion beam to etch three sets of cylinders $\sim 2 \mu\text{m}$ deep into a GaAs/AlGaAs heterostructure through a 2D periodic mask with sub micron pitch. The angle between each set of channels is set at 120° , 35° from the surface. Broadband transmission measurements suggest that these structures do exhibit an optical bandgap near normal incidence that is consistent with model calculations.

Another approach is to use the self-assembly of monodispersed opal¹⁵ or polystyrene¹⁶ spheres with lattice constants from ~ 250 –500 nm. The self assembly is a very attractive attribute of this approach, but it suffers fundamentally from relatively low dielectric contrast. It is necessary to impregnate the voids with an electronically active, high index material to fully exploit them. Progress on these systems is being made, but the stop-bands observed to date are relatively poorly defined, apparently due to lattice imperfections.

Another approach to realize 3D submicron pitch PBMs involves formation of 1D surface gratings on a sacrificial liftoff layer.¹⁷ These textured surfaces are then turned upside down, and fused to a previous layer of rods oriented at 90° . The sacrificial layer is removed, freeing the original substrate, leaving a 1D array of dielectric rods fused to a previously laid-down layer. It is unclear how practical this approach is for reproducibly fabricating high-quality photonic crystals, but it does address the challenging problem of how to achieve high-contrast periodic texture normal to the surface of semiconductor wafers.

Given this challenge in the vertical direction, and given the fact that non-periodic planar dielectric texture is routinely used in optoelectronic circuits to localize photons in slab waveguides, it is natural to explore 2D photonic bandstructures in slab waveguide geometries. An early success in this geometry was reported by Wendt and co-workers.¹⁸ They produced a 2D honeycomb array of holes on a sub 500 nm pitch that penetrated from the surface down ~ 1100 nm into a special slab waveguide fabricated in a GaAs/AlGaAs/InGaAs heterostructure. Photoluminescence excited in the vicinity of this structure showed signs of preferential propagation along certain symmetry direction in the plane of the slab. A much more systematic study of similar structures was recently reported by Labilloy *et al.*,¹⁹ who measured both reflection and transmission of guided photoluminescence through 2D lattices that completely penetrated the slab waveguide.

A group from MIT has concentrated on achieving a higher degree of photon confinement by fabricating a 1D silicon waveguide on an insulating substrate.²⁰ A line of holes was etched through the silicon channel to form in effect a linear, 1D photonic crystal for optical modes confined to the silicon channel. By omitting one of the holes in the center of the array a localized photon was formed with an effective volume of less than $0.1 \mu\text{m}^3$. This localized mode was observed as a

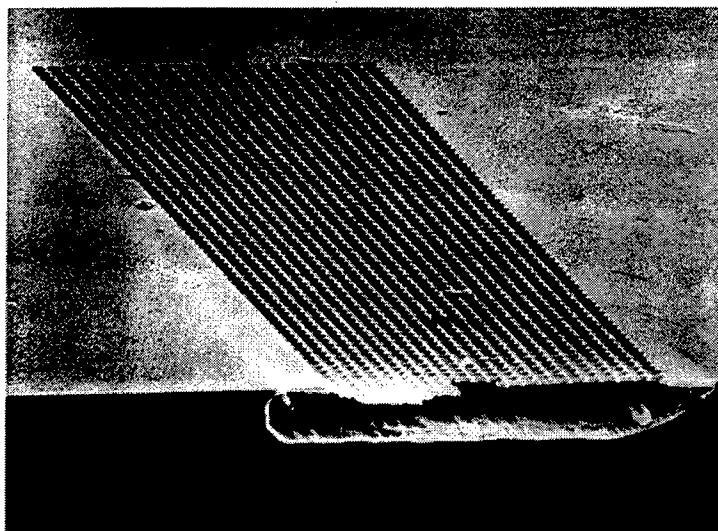


Figure 1. Scanning electron micrograph of a 2D square array of holes that penetrate a 125 nm thick slab of $\text{Al}_{0.15}\text{Ga}_{0.85}\text{As}$. The pitch is 460 nm.

narrow transmission resonance within a stop gap ~ 400 nm wide centered at ~ 1.54 μm .

- *Two-dimensional photonic lattices in high index contrast III-V slabs*

Our group's experimental efforts have concentrated on achieving 2D photonic lattices in similar ultrahigh index contrast structures formed in direct bandgap GaAs/InGaAs/AlGaAs semiconductors rather than silicon. Electron beam lithography is used to define 2D periodic arrays of holes in ~ 200 nm thick PMMA resist layers on GaAs or GaAs/InGaAs epilayers that range in thickness from ~ 70 to 150 nm. These thin "core" layers are separated from the substrate by a sacrificial layer of $\text{Al}_x\text{Ga}_{1-x}\text{As}$ with $x > 0.6$. The resist layer is used as a mask to etch the 2D pattern of holes through the top core layer, into the sacrificial support. The PMMA is then removed and the structure is etched in hydrofluoric acid to remove the $\text{Al}_x\text{Ga}_{1-x}\text{As}$, resulting in an air-bridged porous waveguide (PW) supported from the sides of the patterned region (~ 100 μm wide). Figure 1 shows an SEM micrograph of one such structure after it was cleaved through the processed region. Optical scattering experiments using broadband radiation incident normal to the surface of structures similar to that shown in Fig. 1 have been reported previously.²¹ The following section describes some of the properties of these PW structures in conceptual terms, using model calculations for specific structures as illustrative examples.

3. Optical excitations in periodically textured 2D slab waveguides

It is important to recognize the distinction between a true 2D photonic crystal (like those studied extensively in the microwave regime), and a slab waveguide penetrated by a 2D array of holes. Even if the holes are uniform, the corresponding eigenstates are fundamentally different from those in a true 2D crystal due to the lack of translational invariance along the holes. For instance, the eigenstates with no axial momentum in the true 2D crystal can be labeled by their in-plane wavevector, k_{\parallel} , and their polarization state, either TE or TM (TE and TM polarized states have their electric or magnetic field components parallel to the cylinders respectively). Thus for zero axial momentum, the photonic bandstructure calculation for true 2D crystals reduces to two independent scalar problems, with two independent sets of modes. With a slab geometry the excitations in a 2D crystal can still be labeled by the in-plane wavevectors, but the polarization is quite generally mixed, so the problem can never be rigorously reduced to a scalar one. To appreciate the consequences of this difference, and to illustrate some general properties of 2D photonic crystals in slab waveguides, it is instructive to consider the reflectivity spectra of light incident on them from the vacuum, for all in-plane wavevectors k_{\parallel} within the first Brillouin zone.

Figure 2 shows the calculated reflectivity of *p*-polarized plane waves incident from vacuum on a free-standing infinite GaAs slab textured as described in the caption. The in-plane wavevector of the incident wave is fixed and, for frequencies above the threshold for free propagation in vacuum (0.4 on the x-axis),

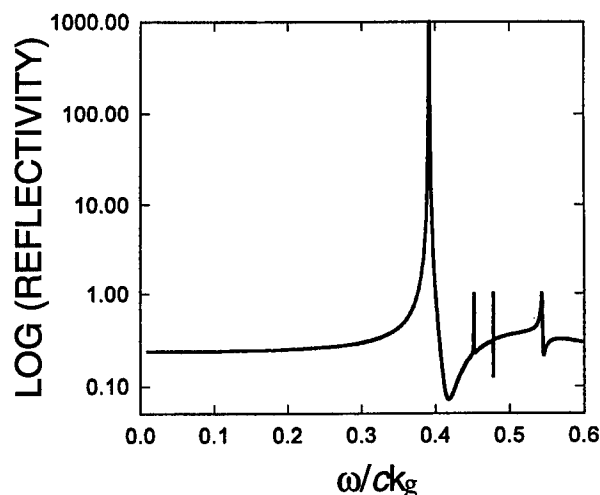


Figure 2. Calculated reflectivity from a 2D square lattice of holes 10 nm deep in a 100 nm thick free-standing GaAs slab. The pitch is 500 nm and the parallel component of the incident wave is held constant at $0.4k_g$ ($k_g = 2\pi/\Lambda$). The plot crosses the light line at 0.4 on the x-axis. The hole filling fraction is 50%.

the reflectivity is, of course, bounded by unity. The Fano-like resonances that peak at unity reflectivity in this frequency range correspond to the excitation of resonant cavity modes of the 2D photonic crystal. They are the leaky remnants of bound slab modes renormalized by multiple scattering from the strong periodic dielectric "potential" associated with the holes. A resonant eigenstate at wavevector k_{\parallel} consists of a quasi-self-sustaining superposition of polarization sheets at wavevectors $k_{\parallel} + (nx + my)k_g$ where $k_g = 2\pi/\Lambda$. Near the center of the Brillouin zone, resonant eigenstates with increasing energy are *dominated* by polarization at wavevectors characterized by increasing values of $\{n, m\}$. The lack of translational invariance perpendicular to the slab in the waveguide geometry means that modes with $\{\omega, k_{\parallel}\}$ above the vacuum light line have a component of polarization at k_{\parallel} that will necessarily radiate into the vacuum. Thus the polarization in the slab is not perfectly self-sustaining, and the modes have finite lifetimes. Another way of describing this effect is to say that the bound modes of the untextured slab become coupled to the vacuum continuum due to the periodic dielectric "potential". The resonant states of the combined system appear in the reflectivity as Fano resonances and their lifetimes are inversely related to the resonance widths.

Away from these Fano resonances, but still above the light line, incident radiation passes through the textured slab without resonantly exciting large internal fields. The corresponding field structures are part of the vacuum excitation continuum that is not significantly modified by the textured slab.

For evanescent fields, incident at frequencies below the light line, the reflectivity is not bounded by unity, and it diverges when resonant with a true bound excitation of the slab (one is evident in Fig. 2, just below the light line). Such excitations *can* exist in the textured slab because none of their polarization components, at $k_{\parallel} + (nx + my)k_g$, is phase matched to radiate into the vacuum. This portion of the excitation spectrum has dispersion characteristics similar to true 2D crystals that are translationally invariant in the perpendicular direction. To a certain extent the 2D dielectric "potential" renormalizes the untextured slab mode dispersion into a photonic bandstructure for (still) bound modes. The difference is that even below the light line, the texture not only couples TE to TE and TM to TM slab modes, but also couples TE and TM modes as it restructures the field envelope perpendicular to the slab.

- *Modeling excitation of 2D porous waveguides*

Figure 2 suggests one way of mapping the dispersion of bound and resonant states of 2D porous waveguides, namely by searching for poles and Fano resonances respectively in the calculated reflectivity as a function of k_{\parallel} . Various techniques can be used to calculate the reflectivity. Full numerical approaches involving numerous basis states and/or grid points are required to accurately model very high index contrast structures such as those shown in Fig. 1. Our group has developed a modified version of Pendry's finite-difference integration scheme²² for this purpose. The modifications are necessary to properly describe the polarization properties of the modes, a quantity that converges very slowly using

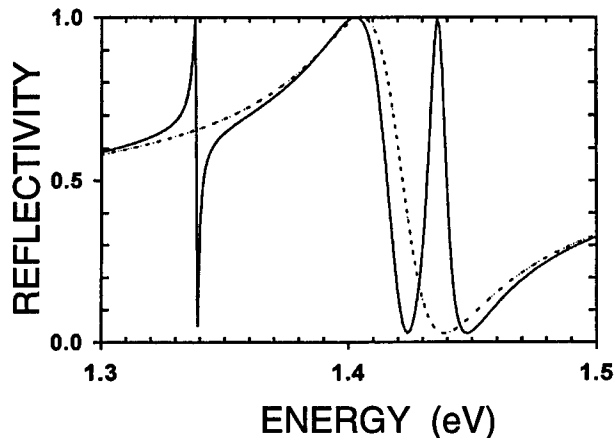


Figure 3. Reflectivity for radiation incident from air on a 50nm film of refractive index $n = 3.5$, textured with a square array of 300 nm diameter holes filled with a material having $n=1.5$, all on a substrate with $n=1.5$. The radiation is incident with in-plane wavevector $k_{||} = 0.08k_0x$. The solid and dashed lines are for s and p polarized radiation respectively.

the original algorithms. Figure 3 shows the Fano resonances obtained from our code applied to a 50 nm thick GaAs slab surrounded by vacuum on the top, and dielectric with $n = 1.5$ on the bottom. Even in the absence of texture, this slab waveguide structure does not support TM modes. The four features evident in the near-normal-incidence reflectivity from the porous structure (see Fig. 3 caption for structural details) correspond to TE-like resonant modes that are characterized by an electric field vector that oscillates *predominantly* in the plane of the waveguide. Each of these "new" eigenstates of the periodically textured 2D slab is composed of a particular superposition of four TE modes associated with wavevectors at $\{\pm k_g + k_{||}, 0\}$ and $\{k_{||}, \pm k_g\}$. Though all four eigenstates are predominantly TE polarized, all also have (small) TM-polarized field components at the above wavevectors. Note that along the $[1,0]$ direction, three of the modes are revealed in the s -polarized reflectivity and one in the p -polarized reflectivity.

The polarization of the radiative component of the resonant states is thus a useful label along symmetry axes. This label is of significant practical relevance since it implies that different resonant states can selectively communicate with the external world through well-defined polarization selection rules.

The numerical schemes for accurately estimating the excitations of high-index-contrast structures tend to be very computationally intensive. To provide intuitive insight into the general symmetry properties of the excitation spectra without having to resort to these time-consuming codes we have developed another method of calculating the reflectivity of 2D-textured waveguides. A Green's function approach is used to calculate the polarization induced in the 2D

textured layer of a slab waveguide. This approach is applicable for large index contrasts, but it is only accurate when the thickness of the textured region is small compared to the wavelength of radiation and the total thickness of the host slab waveguide. The model is described in detail in Ref. 23.

To illustrate some of the unusual properties of these structures, Fig. 4 shows the dispersion of modes that can be excited by using *p*-polarized light to irradiate a 100 nm thick free standing slab of GaAs wherein the top 10% consists of a 2D square array of circular holes. The structure is designed so that the energy of the two TM modes of the untextured guide at the $\{\pm k_g/2, 0\}$ points of the Brillouin zone are nearly degenerate with the four TE modes at $\{\pm k_g/2, \pm k_g\}$. Again, the effect of the texture is to couple these modes together to produce "new" eigenstates in the slab. Zooming in near the zone boundary (see inset of Fig. 4) there are obvious anti-crossings that correspond to the 2D dielectric lattice coupling the TE-like and TM-like slab modes. The inset also shows that modes excited by *s*-polarized radiation do not couple with any of those that are excited by *p*-polarized light. At a given wavevector, the eigenstates in different bands are dominated by different TE and/or TM slab mode components.

4. Discussion and conclusions

The photonic eigenstates in 2D periodically-textured, high-index-contrast slab waveguide structures are qualitatively different from those in true 2D photonic bandgap materials owing to the lack of translational invariance perpendicular to

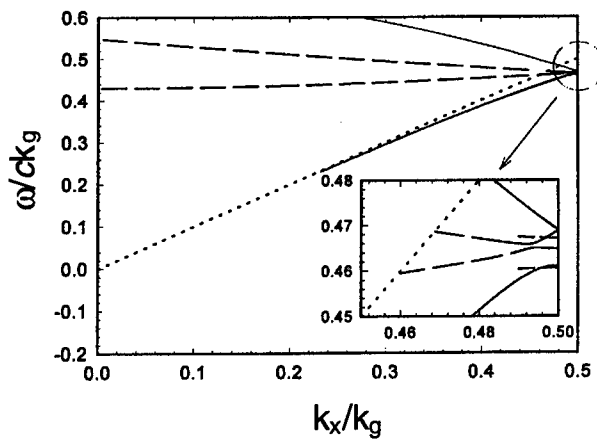


Figure 4. The full dispersion of modes excited by *p*-polarized light calculated along one of the square lattice axes for the structure described in Fig. 2. The dotted curve is the light line, the solid lines are TM-like modes and the dashed lines are TE-like modes. The inset also includes the two modes excited by *s*-polarized light that exhibit nearly horizontal dispersion.

the periodic texture. These waveguide-based structures are easier to fabricate than full 3D photonic lattices in the near-infrared, and it has already been shown that the texture can be used to manipulate their photonic excitation spectra over energies of order 100 meV or more. This capability makes such structures of considerable interest as host materials for future information processing technologies, wherein the photonic eigenstates and their dispersions can be tailored using nanometer-scale patterning processes.

At a given frequency a specific structure will in general support i) bound states with infinite lifetimes; ii) resonant states with finite-lifetimes that are localized in the textured slab but that communicate with the surroundings through the lowest order component of their polarization; and iii) continuum states that are of comparable strength in the slab and the surrounding vacuum or dielectric. By controlling the pitch and filling fraction of the dielectric lattice it is possible to control the relative density of states of these three types of excitations in the frequency range of interest.²⁴ One can imagine composite structures with different regions designed for specific purposes. Resonant states would offer relatively efficient conduits for coupling information into and out of such a structure, whereas the true bound modes would be more natural intermediaries for computational purposes. In dealing with these structures it will be important to recognize and perhaps make effective use of their peculiar polarization properties.

4. Acknowledgements

We wish to thank Simon Watkins for providing us with GaAs/AlGaAs wafers and the Natural Sciences and Engineering Research Council and the Canadian Cable Labs Fund for financial support.

References

1. E. Yablonovitch, "Inhibited spontaneous emission in solid-state physics and electronics," *Phys. Rev. Lett.* **58**, 2059 (1987).
2. S. John, "Strong localization of photons in certain disordered dielectric superlattices," *Phys. Rev. Lett.* **58**, 2486 (1987).
3. W. M. Robertson and G. Arjavalingam, "Measurement of photonic band structure in a two-dimensional periodic dielectric array," *Phys. Rev. Lett.* **68** 2023 (1992).
4. E. Yablonovitch, "Photonic band-gap structures," *J. Opt. Soc. Am. B* **10**, 283 (1993).
5. E. Yablonovitch, T. J. Gmitter, and K. M. Leung, "Photonic band structure: the face-centered-cubic case employing nonspherical atoms," *Phys. Rev. Lett.* **67** 2295 (1991).
6. H. S. Sozuer and J. W. Haus, "Photonic bands: simple-cubic lattice," *J. Opt. Soc. Am. B* **10**, 296 (1993).

7. K. M. Ho, C. T. Chan, and C. M. Soukoulis, "Existence of a photonic gap in periodic dielectric structures," *Phys. Rev. Lett.* **65**, 3152 (1990).
8. J. D. Joannopoulos, R. D. Meade, and J. Winn, *Photonic Crystals*, Princeton, NJ: Princeton University Press, 1995.
9. P. R. Villeneuve and M. Piche, "Photonic band gaps of transverse-electric modes in two-dimensionally periodic media," *J. Opt. Soc. Am. A* **8**, 1296 (1991).
10. S. John and T. Quang, "Localization of superradiance near a photonic bandgap," *Phys. Rev. Lett.* **74**, 3419 (1995).
11. S. John and T. Quang, "Quantum optical spin-glass state of impurity two-level atoms in a photonic band gap," *Phys. Rev. Lett.* **76**, 1320 (1996).
12. C. Weisbuch, M. Nishioka, A. Ishikawa, and Y. Arakawa, "Observation of the coupled exciton-photon mode splitting in a semiconductor quantum microcavity," *Phys. Rev. Lett.* **69**, 3314 (1992).
13. J. J. Baumberg, "Suppressed polariton scattering in semiconductor microcavities," *Phys. Rev. Lett.* **81**, 661 (1998).
14. C. C. Cheng and A. Scherer, "Fabrication of photonic band-gap crystals," *J. Vac. Sci. Technol. B* **13**, 2696 (1995).
15. Y. A. Vlasov, V. N. Astratov, O. Z. Karimov, *et al.*, "Existence of a photonic pseudogap for visible light in synthetic opals," *Phys. Rev. B* **55**, 13357 (1997);
16. R. D. Pradhan, J. A. Bloodgood, and G. H. Watson, "Photonic band structure of bcc colloidal crystals," *Phys. Rev. B* **55**, 9503 (1997).
17. C. Zhang, L. Zavieh, A. Mitra, and T. S. Mayer, "Fabrication of GaAs-based 3-D photonic bandgap materials," *Proc. IEEE Cornell Conf. Adv. Concepts High-Speed Semicond. Dev. Circ.*, Ithaca, NY (1997).
18. J. R. Wendt, G. A. Vawter, P. L. Gourley, T. M. Brennan, and B. E. Hammons, "Nanofabrication of photonic lattice structures in GaAs/AlGaAs," *J. Vac. Sci. Technol. B* **11**, 2637 (1993).
19. D. Labilloy, H. Benisty, C. Weisbuch, *et al.*, "Quantitative measurement of transmission, reflection, and diffraction of two-dimensional photonic band gap structures at near infrared wavelengths," *Phys. Rev. Lett.* **79**, 4147 (1997).
20. J. S. Foresi, P. R. Villeneuve, J. Ferrera, *et al.*, "Photonic-bandgap microcavities in optical waveguides," *Nature* **390**, 143 (1997).
21. M. Kanskar, P. Paddon, V. Pacradouni, *et al.*, "Observation of leaky slab modes in an air-bridged semiconductor waveguide with a two-dimensional photonic lattice," *Appl. Phys. Lett.* **70**, 1438 (1997).
22. P. M. Bell, J. B. Pendry, L. M. Moreno, and A. J. Ward, "A program for calculating photonic band structures and transmission coefficients of complex structures," *Computer Phys. Commun.* **85**, 306 (1995).
23. P. Paddon and J. F. Young, "Simple approach to coupling in textured planar waveguides," *Optics Lett.* **23**, 1529 (1998).
24. S. Fan, P. R. Villeneuve, J. D. Joannopoulos, and E. F. Shubert, "High extraction efficiency of spontaneous emission from slabs of photonic crystals," *Phys. Rev. Lett.* **78**, 3294 (1997).

Intersubband Terahertz Emitters

Q. Hu and B. Xu

Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

M. R. Melloch

School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, U.S.A.

1. Introduction

Terahertz (1–10 THz, or 4–40 meV, or 30–300 μm) frequencies are among the most underdeveloped electromagnetic spectra, even though their potential applications are promising for spectroscopy in chemistry and biology, astrophysics, plasma diagnostics, remote atmospheric sensing and imaging, noninvasive inspection of semiconductor wafers, and communications. This underdevelopment is primarily due to the lack of coherent solid-state THz sources that can provide high radiation intensities (greater than a milliwatt). The THz frequency falls between two other frequency ranges in which conventional semiconductor devices have been well developed. One is the microwave and millimeter-wave frequency range, and the other is the near-infrared and optical frequency range. Semiconductor electronic devices that utilize the classical real-space charge transport (such as transistors, Gunn oscillators, Schottky-diode frequency multipliers, and photomixers) are limited by the transit time and parasitic RC time constants. Consequently, the power level of these classical devices decreases as $1/f^4$, or even faster, as the frequency f increases above 1 THz. Semiconductor photonic devices based on quantum-mechanical interband transitions, however, are limited to frequencies higher than those corresponding to the semiconductor energy gap, which is higher than 10 THz even for narrow-gap lead-salt materials. Thus, the frequency range of 1–10 THz is inaccessible for existing semiconductor devices.

Semiconductor quantum wells are human-made quantum-mechanical systems in which the energy levels can be designed and engineered to be of any value. Consequently, unipolar lasers based on intersubband transitions (electrons that make lasing transitions between subband levels within the conduction band) were proposed for long-wavelength sources as early as the 1970s.¹ More detailed design analyses were reported in the last decade.^{2–8} However, because of the great challenge in epitaxial material growth and the unfavorable fast nonradiative relaxation rate, unipolar intersubband-transition lasers (also called quantum-cascade lasers) at mid-infrared wavelengths (3–5 μm and 8–12 μm) were

developed only recently at Bell Laboratories.^{9,10} This achievement is inspiring, but major obstacles remain for the development of longer wavelength THz intersubband lasers. First, the small energy scales of THz photons make it difficult to detect and analyze spontaneous emission, which is a crucial and necessary step in developing lasers. Second, the energy levels that correspond to THz frequencies are quite narrow (~ 10 meV), so the requirements for the design and fabrication of suitable quantum wells are demanding. Also, for subband separations below one LO-phonon energy (which is 36 meV for GaAs), hot-electron and phonon-bottleneck effects could change the intersubband scattering rates significantly and, therefore, drastically affect population inversion. Third, mode confinement, which is essential for any laser oscillation, is difficult to achieve at THz frequencies. Conventional dielectric-waveguide confinement is not applicable because the evanescent field penetration, which is proportional to the wavelength and is on the order of several tens of microns, is much greater than the active gain medium of several microns.

In this article, we discuss our recent progress towards overcoming the above-mentioned three obstacles. We have constructed a measurement set-up and can routinely gather high-resolution THz emission data. We have designed and fabricated suitable multiple quantum-well (MQW) structures that generate THz spontaneous emission at the designed frequencies. We have performed detailed analysis on gain and losses of THz cavities formed by plasma confinement.

2. Design and measurements of three-level THz emitters

Both intrawell (spatially vertical transition) and interwell (diagonal transition) schemes were utilized in the development of the mid-infrared quantum-cascade lasers¹¹ and were explored in our earlier work on THz emitters.^{6,12} However, our recent success was based on the interwell scheme, mainly because the spatial localization of the two subband wave functions makes selective injection and the removal of electrons easier.^{13,14} This task is difficult for THz lasers, whose subband separation is only ~ 10 meV.

Our MQW structure for THz emission is shown in Fig. 1, in which the conduction band profile and the square of the wave functions were calculated self-consistently from Schrödinger and Poisson equations. The device is formed by a triple-well structure using GaAs/Al_{0.3}Ga_{0.7}As materials, as shown in the dashed box. This structure is essentially a three-level system (marked as E_3 , E_2 , and E_1 in Fig. 1) which is required for any laser. (The level E_4 is much higher in energy so it does not contribute to transport at low bias.) Because there is no recombination involved in unipolar intersubband lasers, electrons can be "reused" many times. Consequently, many identical triple-well modules can be cascade-connected, and the emission power and the mode confinement factor can be increased substantially. Due to translational symmetry, design analysis needs to focus only on one module, provided there are no global space charges and high-field domains. The collector barrier (the one with a 2.0-nm thickness) is center δ -doped

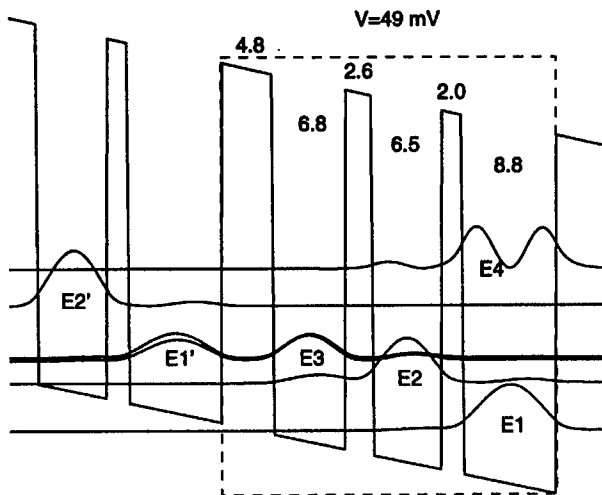


Figure 1. Numerically calculated band diagram, subband levels, and squared magnitude of wave functions of cascade-connected triple quantum-well modules, under a bias of 49 mV/module. The MQWs are made of GaAs/ $\text{Al}_{0.3}\text{Ga}_{0.7}\text{As}$ heterostructures, and the well widths and the barrier thicknesses are indicated in nm. The 2.0 nm barrier is center δ -doped at a level of $1 \times 10^{11}/\text{cm}^2$ in order to provide dynamic charges that assure a global charge neutrality.

at approximately $10^{11}/\text{cm}^2$ in order to provide dynamic charges to assure a global charge neutrality. The radiative transition takes place between E_3 and E_2 , with an energy separation $\Delta E_{32} \approx 15$ meV and an oscillator strength of $f_{32} \approx 0.25$ (using the effective mass in GaAs). Under the designed bias of 49 mV per module, the ground state E_1' of a previous module is aligned with E_3 . Thus, the upper subband E_3 can be selectively populated through resonant tunneling. The energy separation between E_2 and E_1 was designed to be 34 meV under the bias, which is close to the LO-phonon energy $\hbar\omega_{\text{LO}}$ in GaAs. Once energetically allowed, the very fast LO-phonon scattering (with a time $\tau_{21} \approx 1.4$ ps) will rapidly depopulate the E_2 level and establish a population inversion between E_3 and E_2 .

The MQW structures were grown in the molecular-beam epitaxy (MBE) machine at Purdue University. In order to verify the accuracy of our design calculations and to inspect the quality of quantum wells and interfaces, we performed an infrared absorption measurement with the result shown in Fig. 2. The measurement was performed on an 80-module device (with a total of 240 quantum wells) at room temperature. A mid-infrared absorption peak is clearly seen at 110 meV, which is due to the intersubband transition from E_1 to E_4 . The FWHM is only 7 meV, including a 4 meV instrumental linewidth. This narrow linewidth is an indication of the high quality and uniformity of the wells and interfaces. Furthermore, the measured $E_1 \rightarrow E_4$ transition frequency of 110 meV

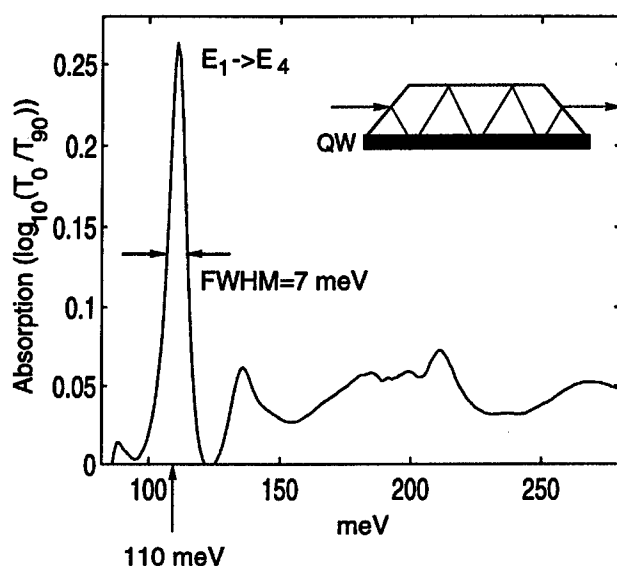


Figure 2. Infrared absorption measurement of a 80-module device, which was placed at room temperature. The absorption peak is due to the $E_1 \rightarrow E_4$ intersubband transition. The measured FWHM is 7 meV, including a 4 meV instrumental linewidth. The measured intersubband transition frequency (110 meV) and dipole moment (14 Å) agreed quite well with the calculated values of 109 meV and 12 Å.

and the dipole moment of 14 Å (deduced from the area of the absorption peak¹⁵) agreed quite well with the calculated values of 109 meV and 12 Å, indicating the accuracy of our calculations.

For intersubband transitions in MQWs, the dipole-interaction selection rule allows only the electric field to be polarized perpendicular to the plane of quantum wells. Consequently, for unpatterned MQWs, edge-coupling is used for intersubband emission. For applications at longer-wavelength THz frequencies, it is desirable to couple the radiation through the surface, from which the beam will be more collimated. Furthermore, cavity loss α tends to be high ($>100 \text{ cm}^{-1}$) at THz frequencies due to free-carrier absorption. Consequently, only photons emitted within $1/\alpha$ from the edge can be coupled out of an edge-emitting structure. For a surface-emitting structure, however, most of the emitted photons can be coupled out of the cavity. Metallic grating coupling is one of the preferred techniques for surface-emitting devices. We have performed numerical analysis for grating couplers for intersubband emission and our results showed that a large coupling efficiency (several tens of percent) is achievable when the grating period is comparable to the photon wavelength.¹⁶

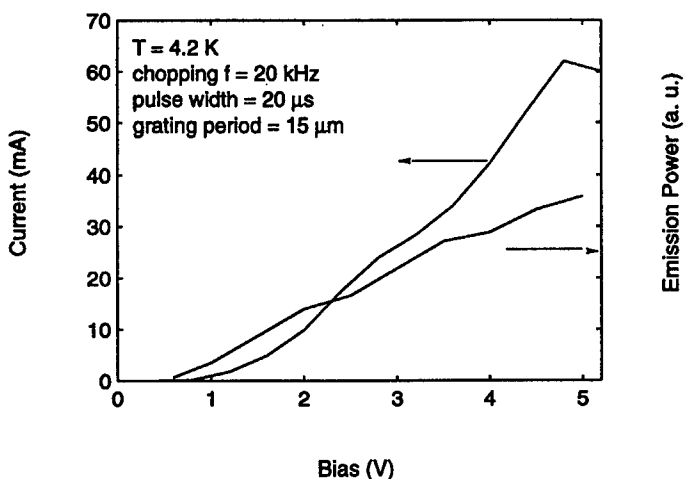


Figure 3. Measured current-voltage (I - V) and emission power-voltage (P - V) curves from a 100-module device in which each module has a triple quantum-well structure shown in Fig. 1.

The measured current-voltage $I(V)$ and emitted power-voltage $P(V)$ curves of a device with 100 triple quantum-well modules are shown in Fig. 3. The current increases monotonically at low biases until the bias voltage reaches the designed value of 4.9 V (49 mV/module), above which the level E_1' becomes misaligned with the level E_3 , and negative dynamic resistance appears. The emitted THz power, detected using a far-infrared detector, increases smoothly in this bias voltage range. This approximately linear $P(V)$ behavior is an indication that each of the 100 modules is sequentially turned on (with a 49 mV voltage drop across the module) as the bias voltage increases.

In order to spectrally resolve the emission signals, we constructed a set-up that included a Fourier transform infrared spectrometer (FTIR) with an external far-infrared detector. The system's schematic is shown in Fig. 4. Using this set-up, we resolved the emission spectra from many different MQW devices and at different biases. A representative one is shown in Fig. 5. This result was obtained from the same 100-module device as shown in Fig. 3 and at a bias of 4.5 V, which is close to the designed value (4.9 V for 100 modules). The center frequency of the spectrum is at 14 meV, which is quite close to the designed value of 15 meV. The FWHM (including a 2-meV instrumental linewidth) is as narrow as 3 meV. The good agreement between the designed and measured emission frequency is strong evidence that our MQW devices function more or less as expected. The narrow emission linewidth shows a high quality of quantum well interfaces and good uniformity among different modules. This narrow linewidth is also important to achieve a high peak gain (which is inversely proportional to the

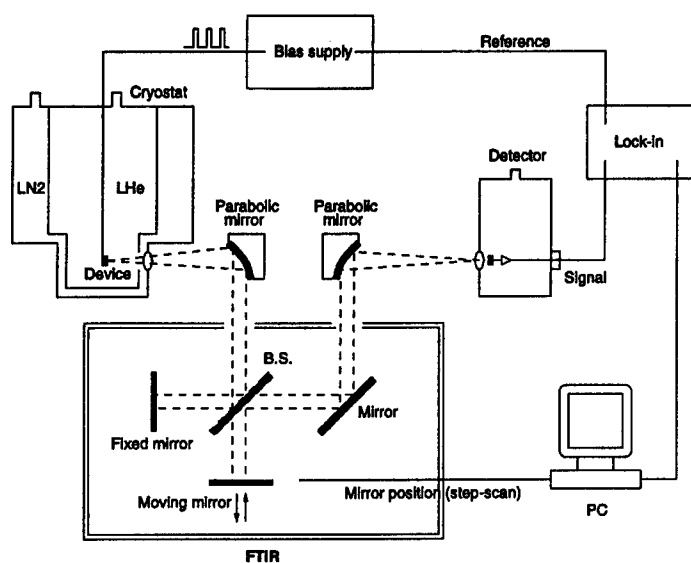


Figure 4. Far-infrared measurement set-up that uses an external Fourier transform spectrometer to spectrally resolve the emitted THz signals.

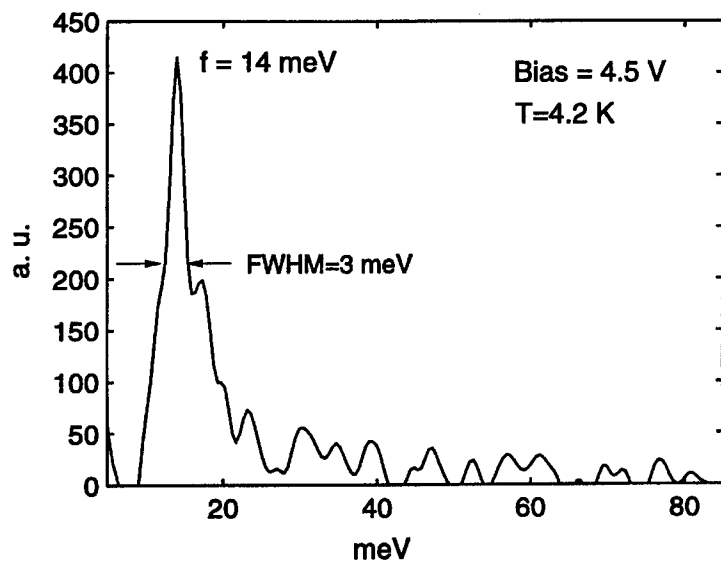


Figure 5. THz emission spectrum from a device with 100 triple quantum-well modules. The measured center frequency is quite close to the design value of 15 meV.

linewidth), that is essential to overcome the inevitable high losses at THz frequencies.

3. Calculated gain and losses of the THz emitters

The intersubband emitters are known to yield a large value of gain, because the two subbands E_3 and E_2 track each other in the k -space; thus electrons emit photons at the same energy regardless of their initial momentum if nonparabolicity is ignored. Therefore, the gain is related to the inverted population density $\Delta n = n_3 - n_2$ in a simple linear fashion, that is,

$$g(\omega) = (\Delta n/t) f_{ij} (e^2/2\epsilon_r^{1/2}\epsilon_0 m^* c) T / [1 + (\omega - \omega_0)^2 T^2] \quad (1)$$

In Eq. (1), t is the thickness of the mode confinement region, and thus $\Delta n/t$ is the three-dimensional inverted population density within t . The quantity $(\pi T)^{-1}$ is the FWHM linewidth of spontaneous emission. Assuming a FWHM of 2 meV (that corresponds to 0.5 THz) and $\Delta n \approx N_D = 10^{11}/\text{cm}^2$, the peak gain at the center frequency is estimated to be approximately 300 cm^{-1} for $f_{32} = 0.31$. This estimate assumes a perfect inverted population, and it will be reduced for a finite n_3/n_2 . (For example, the gain value will be reduced by half for a population ratio of $n_3/n_2 = 3$.) On the other hand, the doping concentration can be made higher than $10^{11}/\text{cm}^2$, resulting in a greater gain.

The population ratio n_3/n_2 is determined by intersubband scattering rates. In a steady state, $n_3/n_2 = \tau_{32}/\tau_{21}$. In order to estimate the population ratio n_3/n_2 , we have calculated various intersubband scattering rates including the effect of temperature, since both the electron and lattice temperatures are likely to be much higher than the ambient temperature when the device is under bias. At an elevated temperature, hot electrons at the high-energy end of the Fermi-Dirac distribution in E_3 can have sufficient energy to emit an LO phonon and be scattered into the lower subband, even though the nominal intersubband spacing is smaller than $\hbar\omega_{\text{LO}}$. Also, at an elevated lattice temperature, the mode occupation number of acoustic phonons can be much greater than unity, thus yielding a shorter scattering time due to the stimulated emission of acoustic phonons. In contrast, the electron-electron intersubband scattering is generally independent of the temperature.¹² Figure 6 shows the $E_3 \rightarrow E_2$ scattering time as a function of the temperature, for an intersubband separation of 16 meV (corresponding to 4 THz). The scattering time due to the LO-phonon emission is obtained by averaging the scattering time over the entire E_3 subband. The scattering time due to the acoustic phonon scattering (including the deformation and piezoelectric processes) includes the stimulated phonon emission process. As shown in Fig. 6, this scattering process is dominated by the LO-phonon emission at temperatures above 50 K, and it can be substantially shorter than its value at low temperatures. In contrast, $\tau_{21} = 1.4 \text{ ps}$ is approximately independent of temperature, since the LO-phonon emission $E_2 \rightarrow E_1$ scattering is energetically allowed even at $T = 0$. Fig. 6 shows that even at a highly elevated temperature of 100 K, $\tau_{32}/\tau_{21} \sim 5$, leading to a large degree of

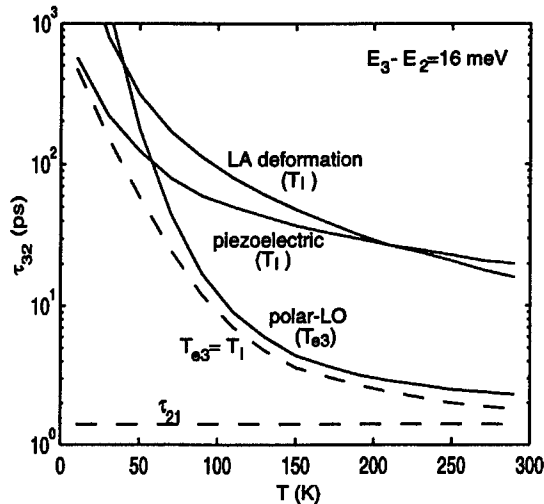


Figure 6. Calculated $E_3 \rightarrow E_2$ scattering times τ_{32} as functions of the temperature. The scattering times due to acoustic phonon emission (both LA and piezoelectric) are only affected by the lattice temperature T_l , while scattering due to LO phonon emission is only affected by the electron temperature T_{e3} . The dashed line shown as $T_{e3} = T_l$ is the total scattering time, assuming $T_{e3} = T_l$. By comparison, the $E_2 \rightarrow E_1$ scattering time (shown as the horizontal line) is independent of temperature, since the LO phonon emission is energetically allowed for the $E_2 \rightarrow E_1$ scattering, even if the electrons are cold.

inverted population $n_3/n_2 \sim 5$. Thus, intersubband THz lasers should be operable up to liquid-nitrogen temperature.

The current density of the device under bias is given by $J = en_3/\tau_3$. From Fig. 6, even at the liquid-nitrogen temperature of 77 K, $\tau_3 \geq 10$ ps, thus yielding a current density of approximately 1 kA/cm² for $n_3 = 10^{11}$ /cm², which is quite easy to handle.

As mentioned in the introduction, one of the most challenging issues in the development of THz lasers is mode confinement, because it cannot be achieved by using conventional dielectric waveguides. Several years ago, one of us proposed the use of plasmas in a heavily doped emitter and collector to confine the emitted THz photons.⁶ Using nonalloyed Ohmic contacts formed on low-temperature grown GaAs, a metal layer can be placed very close to the top side of the MQW structure, as shown in Fig. 7(a). In this scheme, only the bottom side of the device needs the plasma for mode confinement, reducing cavity losses. Fig. 7(b)-(d) show the mode patterns of various field components. Our calculations yielded a mode confinement factor $\Gamma = 0.94$, and a waveguide loss $\alpha = 293$ cm⁻¹. This value is comparable to the available gain, thus lasing based on this scheme is feasible.

In conclusion, intersubband THz emissions at designed frequencies have been observed from MQW devices made of many cascade connected triple-well modules. The observed narrow linewidth (<3 meV) indicates a high quality and

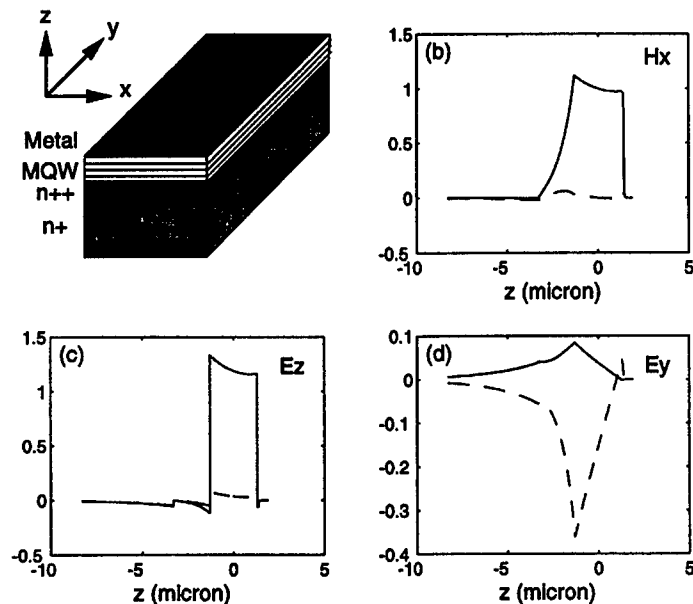


Figure 7. (a) Schematic of a waveguide formed by a metal layer and plasma confinement. (b)-(d) Calculated mode patterns of the TEM mode with the z -direction perpendicular to the QW planes. The dip in E_z corresponds to the $2\text{-}\mu\text{m}$ thick n^{++} plasma region. The active MQW region is $3\text{-}\mu\text{m}$ thick (corresponding to a 100-module device), and the mode confinement factor is $\Gamma = 0.94$.

good uniformity of our MQW structures, and may lead to a high peak gain for THz lasers. Our detailed analysis of gain and losses show that THz lasing is feasible using plasma mode confinement.

4. Acknowledgments

We would like to thank I. Lyubomirsky and B. S. Williams for assistance in the experiments. This work is supported at MIT by the U. S. Army Research Office (DAAH04-95-1-0610) and by NSF MRSEC at Purdue University (DMR-9400415).

References

1. R. F. Kazarinov and R. A. Suris, "Possibility of amplification of electromagnetic waves in a semiconductor with a superlattice," *Sov. Phys. Semicond.* **5**, 707 (1971).

2. F. Capasso, "Band-gap engineering: from physics and materials to new semiconductor devices," *Science* **235**, 172 (1987).
3. P. F. Yuh and K. L. Wang, "Novel infrared band-aligned superlattice laser," *Appl. Phys. Lett.* **51**, 1404 (1987).
4. H. C. Liu, "A novel superlattice infrared source," *J. Appl. Phys.* **63**, 2856 (1988); **69**, 2749 (1991).
5. S. I. Borenstain and J. Katz, "Evaluation of the feasibility of a far-infrared laser based on intersubband transitions in GaAs quantum wells," *Appl. Phys. Lett.* **55**, 654 (1989).
6. Q. Hu and S. Feng, "Feasibility of far-infrared lasers using multiple semiconductor quantum wells," *Appl. Phys. Lett.* **59**, 2923 (1991).
7. A. Kastalsky, V. J. Goldman, and J. H. Abeles, "Possibility of infrared laser in a resonant tunneling structure," *Appl. Phys. Lett.* **59**, 2636 (1991).
8. A. N. Korotkov, D. V. Averin, and K. K. Likharev, "TASERs: Possible dc pumped terahertz lasers using interwell transitions in semiconductor heterostructures," *Appl. Phys. Lett.* **65**, 1865 (1994).
9. J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, and A. Y. Cho, "Quantum cascade laser," *Science* **264**, 477, (1994).
10. C. Sirtori, J. Faist, F. Capasso, D. L. Sivco, A. L. Hutchinson, and A. Y. Cho, "Quantum cascade laser with plasmon-enhanced waveguide operating at 8.4 μm wavelength," *Appl. Phys. Lett.* **66**, 3242 (1995);
C. Sirtori, J. Faist, F. Capasso, D. L. Sivco, A. L. Hutchinson, and A. Y. Cho, "Long wavelength infrared ($\lambda \approx 11 \mu\text{m}$) quantum cascade lasers," *Appl. Phys. Lett.* **69**, 2810 (1996).
11. J. Faist, F. Capasso, C. Sirtori, D. L. Sivco, A. L. Hutchinson, and A. Y. Cho, "Vertical transition quantum cascade laser with Bragg confined excited state," *Appl. Phys. Lett.* **66**, 538 (1995).
12. J. H. Smet, C. G. Fonstad, and Q. Hu, "Intrawell and interwell intersubband transitions in multiple quantum wells for far-infrared sources," *J. Appl. Phys.* **79**, 9305 (1996).
13. B. Xu, Q. Hu, and M. R. Melloch, "Electrically pumped tunable THz emitters based on intersubband transition," *Appl. Phys. Lett.* **71**, 440 (1997).
14. B. Xu, Q. Hu, and M. R. Melloch, "Intersubband THz emission in multiple quantum wells," chapter in: M. Helm, ed., *Long Wavelength Infrared Emitters Based on Quantum Wells*, Vol. 9 of *Optoelectronic Properties of Semiconductors and Superlattices*, New York: Gordon and Breach, 1998.
15. F. Capasso, C. Sirtori, and A. Y. Cho, "Coupled quantum well semiconductors with giant electric field tunable nonlinear optical properties in the infrared," *IEEE J. Quantum Electron.* **30**, 1313 (1994).
16. B. Xu and Q. Hu, "Grating coupling for intersubband emission," *Appl. Phys. Lett.* **70**, 2511 (1997).

Wide-Bandgap Semiconductor Devices for Future Microwave and Millimeter Wave Power Applications

Wallace T. Anderson

Naval Research Laboratory, Washington, DC, 20375

1. Introduction

Present commercially available solid-state devices for high power applications at microwave and mm wave frequencies are limited by the semiconductor materials on which they are fabricated in the amount of power that can be produced per unit area or total perimeter of the device. For example, power densities of GaAs metal electrode semiconductor field effect transistors (MESFETs) and pseudomorphic high electron mobility transistors (PHEMTs) have been reported respectively as 0.40 W/mm and 0.62 W/mm in S-band¹ and 0.35 W/mm and 0.62 W/mm in C-band.² In X-band, the highest power densities we have measured in monolithic microwave integrated circuits (MMICs) delivered to this laboratory under the MIMIC Program have been 0.26 W/mm and 0.32 W/mm for MESFET and PHEMT-based MMICs respectively.

Of the wide-bandgap semiconductors of interest (SiC, GaN, InN, AlN, and diamond),³ SiC^{4,5} and GaN^{6,7} have been the most extensively studied for high power microwave and millimeter wave applications. For the SiC devices, the static induction transistor (SIT)^{8,9} and MESFET¹⁰⁻¹³ are of the most interest for microwave power. SiC heterojunction bipolar transistors (HBTs) and metal oxide semiconductor field effect transistors (MOSFETs) have also been studied, but they are not as promising for microwave power applications because of a number of materials and reliability problems. Although GaN field effect transistors (FETs)^{14,15} have been fabricated and tested, most of the higher frequency wide-bandgap devices are based on heterojunctions of GaN with its alloys. Alloys with GaN have been studied to take advantage of varying the bandgap by changing the composition of the semiconductor material. For example, such "bandgap engineering" has resulted in the fabrication of AlGaIn/GaN modulation-doped HEMTs.¹⁶⁻¹⁹ Other heterojunction devices, such as AlGaIn/GaN HBTs,²⁰ are also under development for future applications.

2. Wide-bandgap microwave and millimeter wave devices

The advantages for microwave and millimeter power applications offered by wide-bandgap semiconductor devices compared to devices based on Si and GaAs have been discussed in many of the above references. These advantages include:

- Much higher breakdown fields allowing higher voltage operation.
- Higher thermal conductivity, particularly in the case of SiC, which is important for removing dissipated heat from power devices.
- Higher temperature operation.
- Higher dc and rf power operation as a result of the much larger current densities that are possible as a result of the higher voltage and higher temperature operation.
- Very high resistivity of semi-insulating (SI) material.

Properties of the wide-bandgap semiconductors of interest are compared with those of Si and GaAs in Table 1 (based on Ref. 3).

Property	Si	GaAs	4H SiC	GaN	AlN	Diamond
Bandgap (eV)	1.1	1.43	3.26	3.45	6.2	5.45
Breakdown Field ($\times 10^5$ V/cm)	3	6	30	> 10	?	100
Thermal conductivity (W/cm \cdot K)	1.5	0.46	4.9	1.3	3.0	22
Resistivity (SI, $\Omega\cdot$ cm)	1,000	10^8	$>10^{12}$	$>10^{10}$	$>10^{13}$	$>10^{13}$
Saturated electron velocity ($\times 10^7$ cm/s)	1.0	1.0	2.0	2.2	?	2.7
Mobility (cm ² /V \cdot s)						
Electron	1,500	8,500	1,140	1,250	?	2,220
Hole	600	400	50	850	?	1,600
Dielectric constant	11.8	12.5	9.6/10	9	8.5	5.5
Lattice constant (Å)	5.43	5.65	3.07	4.51	3.11	3.57
Melting Point (°C)	1,420		2,830			4,000

Table 1. Comparison of wide-bandgap semiconductor properties relevant to microwave and millimeter devices (After Ref. 3).

- *SiC static induction transistors*

Static induction transistors (SITs) are vertical devices in which the current passes from the source on the top of the wafer to the drain of the conducting substrate.⁸ Schottky gates are deposited on the sides of ridges, formed by reactive ion etching (RIE), that support the source contacts. In operation the device is analogous to a triode tube with the depleted SiC between the gate fingers analogous to the vacuum in the tube. Multiple source ridges are fabricated on 2 to 4 μm periods.

In the UHF band, SITs have operated in pulsed mode up to 450 W at 600 MHz.⁸ Transmitter modules have been fabricated with these transistors that operate up to 2.5 kW at 850 MHz.

Recent developmental efforts⁸ have resulted in the highest pulsed power reported to date in L-band: 400 W (100 μs pulses with a 10% duty cycle, 55% efficiency, with 7.7 dB gain, 16.7 W/cm of source periphery) at 1.3 GHz. In S-band airbridges were required and resulted in 78 W (1 μs pulses with a 1% duty, 40 % efficiency, 15.1 W/cm) at 2.9 GHz and 47 W (pulsed, 30% efficiency, 7 dB gain) at 4 GHz.

In UHF, L- and S-bands, SiC SITs will likely provide the largest power as narrow-band discrete devices. Such devices can be readily packaged in high-power modules up to 3 kW. UHF SITs are presently commercially available and L-band SITs should be available by 1999. With sufficient R&D support, S-band SITs should be commercially available at 200 W by the year 2000.

- *SiC MESFETs*

Because they are planar devices, SiC MESFETs may develop somewhat less total power compared with SiC SITs, but they can be fabricated on MMICs, which are useful as microcircuits and for broad-band applications. In a recent developmental effort,¹² SiC MESFETs have achieved a cw power level of 53 W, with 37 % PAE, and a gain of 6 dB when operated at 3.0 GHz with 40 V drain bias. This was a 42 mm periphery device with a 0.7 μm gate length and represents the highest reported cw S-band power achieved to date for a SiC MESFET on a single chip. However, this large device achieved only 1.3 W/mm, while smaller devices (1 mm and below) have been measured at 2.5 W/mm. For example, on wafer measurements at 3 GHz of 1 mm FETs with 55 V on the drain yielded 2.5 W/mm, a PAE of 37 %, with a gain of 12 dB.

The problem with obtaining the expected 2.5 W/mm with the large devices is under investigation. It is thought to be due to trapping (in the active layer, at the active layer/buffer layer interface, and in the SI substrate) and the lack, so far, of a large enough area of uniform material on which to fabricate large devices. The devices were fabricated on 1-3/8 inch diameter SI 4H SiC substrates. For S-band, channel doping was $3 \times 10^{17} \text{ cm}^{-3}$, the gate length L_G was 0.7 μm , the gate width W_G of a single finger was 500 μm , and the gate to drain spacing L_{GD} was 1.5 μm . RF measurements of 1 mm MESFETs, with 10 V on the drain, resulted in a transconductance $g_m = 40 \text{ mS/mm}$, $f_T = 10 \text{ GHz}$, and $f_{\text{max}} = 20 \text{ GHz}$.

An f_T of 10 GHz for a 0.7 μm gate length device is the expected result based on Fig. 1 of Ref. 15. Extrapolated to a 0.1 μm gate length device, which is achieved

commercially today for GaAs pseudomorphic high electron mobility transistors (HEMTs), the f_T is expected to reach 50 GHz. This should allow SiC MESFETs to operate with high power up to 25 GHz, at least into the Ku-band. Support for the possibility of Ku-band operation is provided by the measurement of X-band MESFETs. On wafer measurements of 0.45 μm gate length devices, with $5 \times 10^{17} \text{ cm}^{-3}$ channel doping, resulted in a transconductance of 70 mS/mm, a linear gain of 10.1 to 11.7 dB, and 2.5 W/mm with 41 % PAE at 8 GHz.¹⁰

Addressing the trapping problems and uniformity of material with improved etching techniques and epitaxial layers has resulted in an $L_G = 0.7 \mu\text{m}$ SiC MESFET with cw output power of 80 Watts (1.7 W/mm) in the S-Band (3.0 GHz), with an associated gain of 7.6 dB and a PAE of 38 % for a 48 mm device biased at 58 V.¹³ This is the highest power from a single chip achieved to date in the S-band. The highest power commercial devices (lightly-doped-drain Si MOSFETs) presently produce 60 Watts from a single chip, and this appears to be the upper limit for Si technology. SiC MESFETs should be capable of producing CW power in the range of 120–150 W on a single chip. With adequate R&D support, this could be achieved for commercial devices by 2001.

- *AlGaN/GaN HEMTs*

Besides the advantages of SiC devices, of high-breakdown field (10 times higher than for Si and 5 times than for GaAs), high electron-peak-velocity ($2.2 \times 10^7 \text{ cm/s}$ for GaN), and high temperature applications (low minority carrier thermal-generation rate), GaN, AlN, InN, and their alloys form heterostructures which allow the fabrication of devices such as AlGaN/GaN HEMTs. Compared to conventional MESFETs, HEMTs have higher mobilities, better charge confinement, and higher gate breakdown voltages.

HEMTs have been fabricated on 4H SI SiC substrates by epitaxial growth of an AlN buffer layer followed by a 2 nm layer of undoped GaN followed by a 27 nm layer of $\text{Al}_{0.14}\text{Ga}_{0.86}\text{N}$.¹⁸ A cap layer followed consisting of a 5 nm AlGaN undoped spacer layer followed by a 12 nm donor layer and a 10 nm undoped barrier layer. These HEMTs had $L_G = 0.45 \mu\text{m}$, $L_{GS} = 1.0 \mu\text{m}$, and $L_{GD} = 1.5 \mu\text{m}$, resulting in a transconductance of 200 mS/mm. At 10 GHz a very high power density of 5.28 W/mm was measured, with a PAE of 35.9% and associated gain of 9.17 dB. At the bias point of $V_{DS} = 20 \text{ V}$, $V_{GS} = -1 \text{ V}$, f_T was 28 GHz and f_{max} was 114 GHz.

While SiC may be the best substrate for technical reasons, there are affordability considerations⁷ for growing on sapphire. HEMTs were fabricated on sapphire¹⁹ by first growing a thick GaN buffer layer followed by a strained layer of AlGaN. This strained layer produces free carriers in the two dimensional electron gas (2DEG) in the GaN as a result of the piezoelectric effect. This approach requires no dopant atoms, so the 2DEG carrier concentration across a wafer may be more uniform. The resulting HEMTs, fabricated with 0.15 μm gates, had the best rf performance reported to date, with $f_T = 68 \text{ GHz}$ and $f_{max} = 140 \text{ GHz}$.

Heat removal is one of the major problems, particularly when devices are fabricated on sapphire substrates. SiC has a much higher thermal conductivity than sapphire and is a better lattice match to GaN, but only small wafers (1.5 inch

diameter) are available today, and at high cost. Flip-chip is the most likely solution for heat removal for GaN devices on sapphire.

One of the major problems with the development of GaN devices is that there is as yet no lattice matched substrate on which to grow the material. Epitaxial films grown on SiC have 10^8 defects/cm², while films on sapphire have 10^9 defects/cm². A promising new method of lateral epitaxial overgrowth (LEO) is producing essentially defect free GaN layers on SiC substrates in the laboratory.²¹ AlGaIn/GaN HEMTs fabricated on these layers are expected to have much enhanced high frequency and high power performance compared to similar devices fabricated on sapphire substrates.

Based on the progress to date and research planned to reduce the defect density, AlGaIn/GaN HEMTs should find important future applications in X-, Ku-, and Ka-bands up to 40 GHz. A cw power level of 10 W on a single chip should be possible at 40 GHz with AlGaIn/GaN HEMTs fabricated on SiC substrates. However, GaN devices will require more time to develop than SiC technology, so AlGaIn/GaN-based MMICs may not reach commercial availability until 2005.

3. Wide-bandgap semiconductor device reliability

Despite considerable research effort, currently available microwave Si and GaAs solid-state three-terminal power devices, and MMICs based on them, still have problems to overcome in the areas of reliability, affordability, insufficient PAE and power output. Both GaAs-based power PHEMTs and power HBTs exhibit reliability problems that keep them from being fully utilized. Although the DARPA MAFET program is addressing some of the reliability issues, specific gaps will continue to exist, as they did during the duration of the MIMIC program.²² The performance/cost/reliability potential of devices based on SiC and GaN wide-bandgap semiconductors is of great interest in overcoming the present limitations. Currently, the database on SiC and GaN based rf power device reliability is essentially nonexistent, with no significant studies reported to date.

The objectives of wide-bandgap semiconductor device reliability studies should be to establish the failure mechanisms (for example, by accelerated stress testing), to use advanced diagnostic techniques to determine failure mechanisms, and to interpret the results using model-assisted analysis. In a comprehensive reliability study of wide-bandgap semiconductor devices, a number of possible failure mechanisms should be considered, based on previous work on compound semiconductor devices:

- Ohmic contact degradation can occur by interdiffusion of the contact metal with the semiconductor, resulting in a decrease in the *n*- or *p*- type doping at the interface. A high carrier concentration must be maintained at the interface to allow low resistance tunneling contacts to remain unchanged during the lifetime of the device. An example would be the Ni/SiC ohmic contacts for SiC MESFETs. Ion implantation is a new method used to increase the carrier concentration at the interface. The resistance of this contact will increase

during the lifetime of the device if there is any significant interdiffusion at the interface resulting in a decrease in the carrier concentration.

- Schottky barrier gates can become leaky and lose their sharp rectifying characteristics if there is interdiffusion at the metal/semiconductor interface. This is particularly important with wide-bandgap devices that operate at much higher voltages and electric fields compared to previous technologies. An example would be the Ni gates to SiC MESFETs, which can have 40 V swings on the input RF signal at the gate while already maintaining a 40 V bias on the drain. This has been observed during die attachment of SiC chips to Cu substrates using AuGe eutectic preforms, which melt at 356 °C.
- Doping in the channel can change during the lifetime of a device, resulting in a decrease in gain and loss of RF power and efficiency. An example would be AlGaIn/GaN HEMTs on SiC substrates. In many of these devices the primary carrier concentration in the 2DEG is induced by the piezo-electric effect resulting from the strain in the AlGaIn layer. If the strain is reduced during the lifetime of the device, e.g., by the creation of defects in the LEO material, device performance will degrade.
- In MMICs fabricated with SiC MESFETs or AlGaIn/GaN HEMTs, new types of on-chip capacitors will need to be developed that can maintain 80 to 100 V, much higher than present technology. Defects in the dielectric layer in MIM capacitors can result in failure of the MMIC by burn-out of a capacitor. This is already a problem with high power GaAs MMICs.
- With all the devices (SiC SITs and MESFETs and AlGaIn HEMTs, and MMICs based on these devices) passivation layers are essential for long term reliability. Two common passivations are SiO₂ and silicon nitride, both of which have a piezo-electric effect, and therefore induce charge in the semiconductor, at the passivation/semiconductor interface as a result in the strain induced in the semiconductor. This can be a particular problem with the AlGaIn HEMTs if this strain changes during the life time of the device. Passivation failure is a common failure mechanism with GaAs MMICs and similar problems are expected with the wide-bandgap devices.

4. Conclusions — future applications

Wide-bandgap semiconductor devices hold the best promise for obtaining the highest power from solid-state devices in the 0.5–40 GHz frequency range. In the UHF, L-, and S-bands, SiC static induction transistors (SITs) will likely provide the largest power as narrow band discrete devices, while SiC MESFET-based MMICs should provide the highest broad band power. For S-, C-, and X-bands SiC MESFETs and MESFET-based MMICs will likely provide the highest power in narrow and wide-band applications. SiC MESFET technology may also prove useful into Ku-band. In higher bands, from X-, Ku-, and Ka-bands up to 40 GHz, AlGaIn/GaN HEMTs and HEMT-based MMICs are the best candidates for both

narrow-band and wide-band high power applications. A cw power level of 10 W on a single chip at 40 GHz should be achievable using AlGaIn/GaN HEMTs fabricated on SiC substrates. The successful development of these devices depends on the solution of many problems in the areas of materials, device and circuit modelling and design, packaging, and affordability.

The future time at which the emerging wide-bandgap technology will be commercially available depends not only on the technical problems still to be overcome, but also on the level of support for the research and development effort. Presently, only UHF SiC SITs are sold commercially, with a small number of L-band SiC FETs sold on special order. With sufficient support, S-band SITs should be commercially available at 200 W by the year 2000, and S- and X-band 100 W SiC MESFETs by 2001. SiC MESFET based MMICs should be available by 2002. GaN devices will require more time to develop and may be available by 2005. However, previous work on SiC devices could shorten this period. The development of SiC MESFET based MMICs should act as a stepping stone for the later development of GaN HEMT based MMICs which will operate with higher gain and at higher frequencies.

5. Acknowledgment

This work was supported by the Office of Naval Research.

References

1. L. Aucoin, S. Bouthillette, A. Platzker, *et al.*, "Large periphery, high power pseudomorphic HEMTs," in: *Tech. Digest 1993 GaAs IC Symp.*, Parsippany, NJ: IEEE, 1993, p. 351.
2. S. T. Fu, J. J. Komiak, L. F. Lester, *et al.*, "C-band 20 watt internally matched GaAs-based pseudomorphic HEMT power amplifiers," *Tech. Digest 1993 GaAs IC Symp.*, Parsippany, NJ: IEEE, 1993, p. 355.
3. M. N. Yoder, "Wide bandgap semiconductor materials and devices," *IEEE Trans. Electron Dev.* **43**, 1633 (1996).
4. J. B. Casady and R. W. Johnson, "Status of silicon carbide (SiC) as a wide-bandgap semiconductor for high-temperature applications: a review," *Solid State Electronics* **39**, 1409 (1996).
5. C. Weitzel, "Silicon carbide high-frequency devices," in: *Proc. Intern. Conf. Silicon Carbide, III-Nitrides Related Mater.*, Linkoping, Sweden, 1977, p. 172.
6. J. C. Zolper, J. Jun, T. Suski, J. M. Baranowski, and S. B. Van Deusen, "Advanced processing of GaN for electronic devices: progress and prospects," in: *Proc. 1997 Intern. Semicond. Dev. Res. Symp.*, Charlottesville, VA, 1997, p. 543.
7. M. Razeghi, "21st century: the final frontier for III-nitride materials and devices," in this book.

8. R. J. Bojko, R. R. Siergiej, G. W. Eldridge, *et al.*, "Recent progress in 4H-SiC static induction transistors for high-frequency power generation," *DRC Tech. Digest* (1998), p. 96.
9. A. K. Agarwal, L. S. Chen, G. W. Eldridge, R. R. Siergiej, and R. C. Clarke, "Ion-implanted static induction transistors in 4H-SiC," *DRC Tech. Digest* (1998), p. 94.
10. S. T. Allen, R. A. Sadler, T. S. Alcorn, J. W. Palmour, and C. H. Carter, Jr., "Silicon carbide MESFETs for high-power S-Band applications," in: *Proc. Intern. Conf. Silicon Carbide, III-Nitrides Related Mater.*, Linköping, Sweden, 1977, p. 176.
11. O. Noblanc, C. Arnod, E. Chartier, and C. K. Brylinski, "Characterization of power MESFETs on 4H-SiC conductive and semi-insulating wafers," in: *Proc. Intern. Conf. Silicon Carbide, III-Nitrides Related Mater.*, Linköping, Sweden, 1977, p. 174.
12. R. A. Sadler, S. T. Allen, T. S. Alcorn, *et al.*, "SiC MESFET with output power of 50 W cw at S-Band," *DRC Tech. Digest* (1998), p. 92.
13. S. T. Allen and J. W. Palmour, private communication from Cree Research, Inc., August 1998.
14. S. C. Binari, K. Doverspike, G. Kelner, H. B. Dietrich, and A. E. Wickenden, "GaN FETs for microwave and high-temperature applications," *Solid State Electronics* **41**, 177 (1997).
15. S. C. Binari, W. Kruppa, H. B. Dietrich, *et al.*, "Fabrication and characterization of GaN FETs," *Solid State Electronics* **41**, 1549 (1997).
16. S. C. Binari, J. M. Redwing, G. Kelner, and W. Kruppa, "AlGaIn/GaN HEMTs grown on SiC substrates," *Electron. Lett.* **33**, 242 (1997).
17. M. A. Sánchez-García, E. Calleja, F. Calle, F. J. Sanchez, and E. Muñoz, "Properties of III-nitrides grown on Si(111) substrates by plasma-assisted molecular beam epitaxy," in this book.
18. S. T. Sheppard, K. Doverspike, W. L. Pribble, *et al.*, "High-power microwave GaN/AlGaIn HEMTs on SiC," in: *Proc. 1997 Intern. Semicond. Dev. Res. Symp.*, Charlottesville, VA, 1998, late news paper.
19. K. K. Chu, J. A. Smart, J. R. Shealy, and L. F. Eastman, "AlGaIn/GaN piezoelectric HEMTs with submicron gates on sapphire," to appear in: *Proc. Electrochem. Soc. Meeting*, Boston, MA, 1998.
20. L. S. McCarthy, P. Kozodoy, S. P. DenBaars, M. Rodwell, and U. K. Mishra, "First demonstration of an AlGaIn/GaN heterojunction bipolar transistor," *25th Intern. Symp. Compound Semicond.*, Nara, Japan, October 1998.
21. P. Fini, L. Zhao, J. S. Speck, and S. P. DenBaars, "Selective area growth and lateral epitaxial growth of MOCVD GaN on GaN nucleation layers," to appear in: *Proc. 25th Intern. Symp. Compound Semicond.*, Nara, Japan, 1998.
22. W. T. Anderson, J. A. Roussos, and K. A. Christianson, "MIMIC power amplifier reliability studies," in: *Tech. Digest 1994 Government Microcircuit Applications Conf.*, New York: Palisades Institute for Research Services, 1994, pp. 443-446.

Polymer Optical Interconnects

L. Eldada

Advanced Technologies, AlliedSignal Inc., Morristown, NJ 07962, U.S.A.

1. Introduction

The computer industry has based its systems on metal-interconnected silicon chips for decades, relying on improvements in photolithographic resolution to increase circuit density and computer speed. Efforts in microelectronics have been focusing on scaling down transistor dimensions to decrease their switching delay, although the speed bottleneck does not reside in transistors but in interconnects. The fastest transistors today have a switching delay that is two orders of magnitude smaller than that of the fastest switching design in a computer. We could therefore have, even today, machines that are faster by this factor if we could solve the interconnect problem. Optical interconnect technologies geared toward massively parallel processing applications have come a long way in recent years. Accelerated by several research programs focused on this area, recent progress has shown that optical interconnects are not merely a good idea in theory, but a viable solution to the interconnect bottleneck. In this article, we describe a robust and versatile planar polymer waveguide technology for low-cost, high-performance optical interconnects that address the needs of the computer industry. We also discuss the integration of polymeric optical interconnects on chips, multi-chip modules (MCMs), boards, and backplanes. Light coupling to and from optoelectronic chips is achieved by patterning mirrors directly onto waveguides. We further discuss the high-speed low-cost manufacturing of polymer optical interconnect circuits in an industrial environment.

2. Background

Improvements in the critical dimension of devices and increases in computer speeds have been following Moore's law for decades (Fig. 1), making it seem that they will continue to do so indefinitely. But an objective evaluation of the current barriers indicates that this progress is bound to slow down, as the brute force approach of using the same material system and simply shrinking dimensions is quickly reaching its fundamental limits, with today's 193 nm optical lithography becoming insufficient for the 0.10 μm resolution needed around the year 2006. Lithographic processes using alternative sources, such as extreme ultraviolet (EUV) and x-rays, are not progressing as fast as the requirements for them are advancing, while electron beam patterning remains extremely slow. Furthermore, little room is left for increased density through invention, now that nearly all the

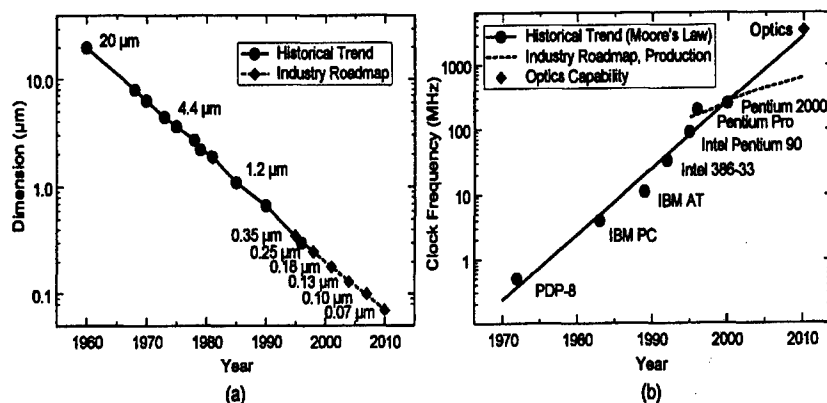


Figure 1. Progress of (a) critical dimension and (b) clock frequency with time. Optics can permit the computer industry to sustain historical exponential growth.

unused space has been removed from semiconductor dice. This effect has been causing the slope of the complexity curve in Moore's law to become shallower. Further aggravated by the recent surge in demand for bandwidth, the situation calls for a new paradigm.

Higher computation speed pushes signal propagation into the realm of optics (Fig. 1(b)). Optical interconnects allow to tap into the tremendous bandwidth of optics while offering revolutionary possibilities in computer architectures.

3. Optical vs. electrical interconnects

The exponential growth in the electronic VLSI technology has been achieved by both decreasing the minimum device feature size and increasing the chip size. But as transistor dimensions are scaled down, the switching delay of the transistor decreases by the same factor as the interconnect resistance per unit length increases. As a result, the RC time constant and the interconnect delay remain unchanged, thereby imposing a limit on the chip speed. And as the chip size is scaled up, increasing the number of elements N contained in a chip, the number of links needed for interconnects between chips increases as $N^{2/3}$, while the available chip perimeter grows only as $N^{1/2}$. Also, since electrical interconnects cannot cross, they must be routed over or under one another through multiple interconnect layers. The length of interconnects and the fan-out number grow with machine complexity, causing a capacitive loading problem for electrical interconnects.¹ Input/output (I/O) circuits drive the capacitance of the devices attached to the gate in addition to the capacitance of the interconnect lines themselves, which is proportional to the length of the lines. Electrical I/O circuitry can consume up to 80% of the total power.² Furthermore, there are parasitic reactance and impedance mismatch problems related to electrical interconnects. Processor speeds have

increased significantly, but computer speed have not increased at the same rate due to the bottleneck in metallic interconnects.

Optical interconnects can be used at all levels in digital computers including cabinet to cabinet, backplane, board to board, MCM to MCM, chip to chip, and intrachip. For communication between cabinets, on backplanes, and between boards, the distances are kilometers to centimeters. These optical links can be built with commercially available optical sources, detectors, fibers and waveguides. Cabinet to cabinet interconnects are in the realm of fiber optic communications, but technologies cannot be transferred directly. Computing requirements are more stringent in terms of cost, power dissipation, accuracy, and tolerable error rates. Also, there are compatibility requirements with CMOS technology and packaging, and differences in multiplexing and bandwidth needs. Board-to-board communications can use optical fibers, planar guides, as well as free-space optics. For communication between MCMs, between chips, and on the intrachip level, the distances are centimeters to microns, the density is high, and the level of flexibility required is high. Optical interconnect features that are especially useful at this level include freedom from mutual coupling effects and flexible routing through three-dimensional space. Challenges for optical chip-to-chip interconnects are mainly in packaging, as it relates to optical alignment, the presence of both GaAs (for optoelectronic components) and Si (for electronic components), and the limited space above the chip plane. In terms of the speed limits of optical interconnects, the data rate is limited by the speed of optoelectronic transceivers, modulators and the interfaces. Free space interconnects are limited by diffraction.

Table 1 shows how electrical and optical interconnects satisfy various criteria. Optical interconnects provide advantages with regard to all criteria and at all levels except on the intrachip level (e.g., gate to gate), because the power efficiency of the electric-optical-electric conversion is still too low today. This makes optical interconnects noncompetitive at very short distances, where electrical interconnect lines have their lowest power consumption because they need no termination when they are shorter than the travel distance of the signal in a bit period.³ But when data leave the chip, optical interconnects become very promising, especially in terms of power consumption, density, data rate, crosstalk, fan-out, and the overall weight, simplicity, cost, and system reliability.

4. Optical interconnects

Optical interconnects provided the infrastructure critical to the explosive growth of telecom that followed, and paved the way for today's flourishing information industry. The technical drivers for the replacement of copper wires with optical fibers include wider data bandwidths, reduced noise, lower loss, and reduced communication cost in \$/bit/km. The adoption of telecom optical interconnects by computer designers has been slow, with the main reasons being the high cost of optoelectronic components and module assemblies, the lack of a transmission

Criteria	E/I	O/I	Reasons
High spatial density High data rates		√	Interaction of electrons is proportional to signal frequency. Photons do not interact with other photons. O/I support larger bandwidth. Impedance mismatch is a limiting factor in E/I. O/I offer impedance matching (photodetectors are quantum detectors, optical transmission lines do not need termination).
Low speed Short distance	√		For short distances, O/I require more power due to low power efficiency in optic-electric conversion of source and detector. Cost/performance benefit does not justify O/I below 1 GHz.
High speed Long distance		√	E/I require more power (large capacitance of bonding pads). Lines are terminated if longer than travel distance of signal in one bit period to eliminate reflection, increasing drive power.
Large fanout		√	Electronic chips cannot function at full speed and produce required current when fanout is high. Lasers deliver ~10 dBm, receiver sensitivity is ~-40 dBm, allowing large optical fanout.
Power consumption		√	Power grows much faster for E/I. As distance increases, ratio of power dissipation to data rate is smaller in O/I.
Routing flexibility		√	Free-space O/I, fibers, and planar strips offer flexibility. Beams share paths without interacting.
Dynamic reconfiguration		√	Free-space O/I have no mechanical contacts. Guided O/I use dynamic holograms with photorefractive materials.
EMI, ground loops		√	O/I are immune to interference and ground loops.
Crosstalk		√	O/I are immune to crosstalk.
Parallelism		√	O/I offer inherent parallelism.
3D routing		√	O/I can exploit the third dimension.
Skew		√	O/I offer less clock and signal skew.
Simplicity		√	O/I require simpler hardware. O/I lend themselves to
Weight		√	multiplexing and switching, permitting less links.
Reliability		√	Failure rate in high-pin-count chips connected electrically renders system insufficiently reliable.
Cost		√	O/I offer larger bandwidth, larger fanout, less links, less power consumption, and simpler hardware.

Table 1. Comparison of electrical and optical interconnects (E/I vs. O/I).

medium that can be planar processed for board-level applications, and the lack of a packaging platform that can be used for both optical and electronic devices. The increased complexity required for electrical wiring solutions for high-speed, high-density, cross-platform interconnects has made such systems quite costly. Recent technological advances have imposed new demands on interconnect performance. These advances include increased chip speeds coming from smaller transistors, increased interconnect complexity, pin count, and processing power (with clock speeds now at 500 MHz and increasing); considerable increase in the number of processors in parallel processing supercomputers; and the exponential growth of network computing and global communication networks (e.g., Internet) that demands wideband interconnects in LANs and WANs. A cost-effective, wideband interconnect solution for high speed data communications between chips, MCMs, boards, backplanes, cabinets, and processors is critically needed in all these systems. Today, data transfer rates in Gb/s are common in distributed computing

	CMOS	VCSEL
Wafer processing		
- Round wafer	✓	✓
- 1D, 2D arrays	✓	✓
- Planar interconnects	✓	✓
- Batch fabrication	✓	✓
- Thin film metallization	✓	✓
- Standard voltage, current	✓	✓
- Standard IC equipment	✓	✓
- Wafer-level tests	✓	✓
- Reliability, stability	✓	✓
- High speed, low threshold	✓	✓
First-level packaging		
- Interconnect methods	WB, TAB, SMT, flip-chip	WB, TAB, SMT, flip-chip
- Substrates	DIP, PGA, BGA, LF	DIP, PGA, BGA, LF
- Alignment	N/A	passive

Table 2. Comparison of wafer processing and first-level packaging in CMOS and VCSEL technologies.

systems, interactive multimedia, high-speed ATM switching, and network communication. High-speed optical interconnects are needed to remove the interconnect bottlenecks among multiple processors in a computer system.

Recent advances that made optical interconnects the wideband technology of choice include VCSEL (vertical cavity surface emitting laser) devices, low-cost VCSEL packaging, polymer-based optical waveguide and fiber materials, low-cost micro-optical components, passive alignment, low-cost connectors, and low-cost fabrication processes (IC-like planar batch processing), allowing low-cost volume manufacturing. Table 2 compares wafer-level processing and first-level packaging for CMOS and VCSEL technologies, revealing a high level of similarity and compatibility.

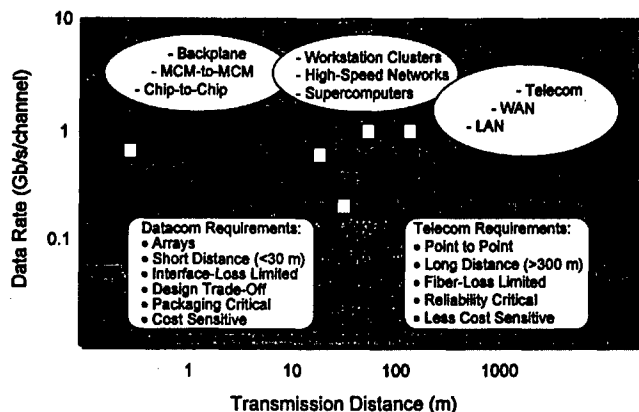


Figure 2. Optoelectronic interconnect development programs in the U.S. and the targeted application areas. The insets illustrate the differences in requirements for optical interconnect technologies used in datacom and telecom systems.

In recent years, several research programs in the U.S. have focused on developing viable optical interconnect technologies for datacom applications. Figure 2 shows graphically the areas addressed by five major programs in terms of data rate and transmission distance.⁴ The POINT (Polymer Optical Interconnect Technology) program is the one relevant to short-distance communication of less than a few meters (e.g., within a cabinet).

5. Polymeric optical interconnects

A number of optical interconnect technologies exist today, but they not all are appropriate for low-cost markets. Figure 3 summarizes several industry studies that describe the economic drivers in the optical interconnect market. The dashed lines represent industry needs at different points in time and the solid lines represent the cost per port as a function of the port count in a component for three key technologies, namely fused fiber, planar glass, and planar polymer. The cost per port in all technologies decreases slowly with the number of ports, then rises sharply when the limits of the technology are reached. The 1993 dashed line shows that 5 years ago, the industry was willing to pay \$50-75 per port, with little need for more than 16 port components. At that point, all three technologies were able to satisfy the customer while being profitable. Today, customers would like the cost per port to be no more than \$15-25, a price that fused fiber components cannot meet, and they need devices with up to 32 ports, a difficult task to achieve with glass fiber. It is expected that, in half a decade, the customer will be willing to pay only \$5-7 per port for components with up to 128 ports. At that point, optical polymers will be the key player. This analysis applies qualitatively to most types of optical components, but quantitatively only to simple devices such as power splitters.

Several polymeric systems are available on the market today. A comparison of the key properties of optical polymers investigated under the POINT program

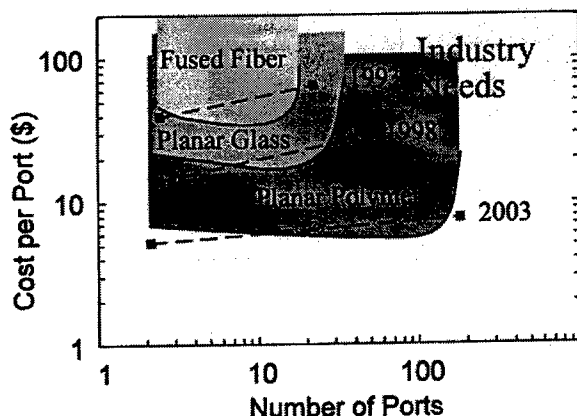


Figure 3. Fundamental economic drivers of optical interconnect circuitry.

can be found in Ref. 4. At AlliedSignal, an advanced polymeric waveguide technology was developed for affordable planar optical interconnects that address the needs of the computer industry.⁵ High-performance organic polymers that can be readily processed into optical waveguide structures of controlled geometries and numerical apertures have been developed. These materials are formed from highly crosslinked acrylate monomers with specific linkages that determine flexibility, toughness, loss, and environmental stability. These monomers are intermiscible, providing for adjustment of the refractive index from 1.3 to 1.6 with 10^{-4} accuracy. In polymer form, they exhibit state-of-the-art loss values (0.02 and 0.001 dB/cm at 840 nm for acrylates and halogenated acrylates, respectively), high thermal stability (65 years at 100 °C induce 0.1 dB/cm loss), and high humidity resistance (no effect on guide transmission after 600 hours at 85 °C and 85% relative humidity).

Waveguides are formed photolithographically, with the liquid monomer mixture polymerizing upon UV illumination either in a mask exposure or by direct laser writing. A wide range of rigid and flexible substrates can be used, including glass, quartz, silicon, glass-filled epoxy, and flexible plastic films.

Laser-delineated multimode waveguides are depicted in Fig. 4. One of the guides (a) was terminated by cleaving the silicon substrate on which it was fabricated and the other (b) has a 90° facet obtained by direct laser termination. The high contrast of the materials results in sharp vertical walls, even in structures that are several hundred microns thick, making it possible to print micro-optical elements.⁶ Another useful aspect of these materials is that, due to the nature of the lithographic process, selective undercutting can be used to make structures that can grip optical fibers (Fig. 4(c)), resulting in a simple and inexpensive fiber pigtailing process. Fibers can be "snapped" into these grippers after development, when the crosslinked polymer is highly elastomeric. This technique is completely passive, making it possible to pigtail optical waveguides rapidly and inexpensively. Printing of fiber grippers and guides is achieved in the same photoexposure step, costing no additional time or effort, thus lending itself well to manufacturing.

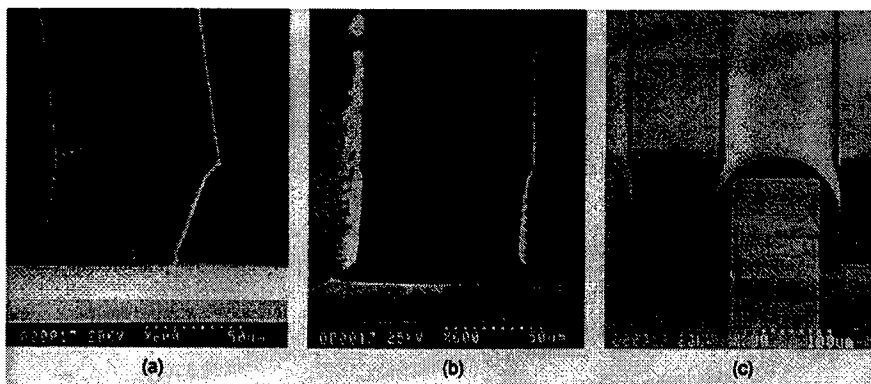


Figure 4. Laser-fabricated multimode optical waveguides terminated (a) by cleaving the silicon substrate and (b) by direct laser termination. (c) A device pigtailed using AlliedSignal's optical polymer fiber gripper technology.

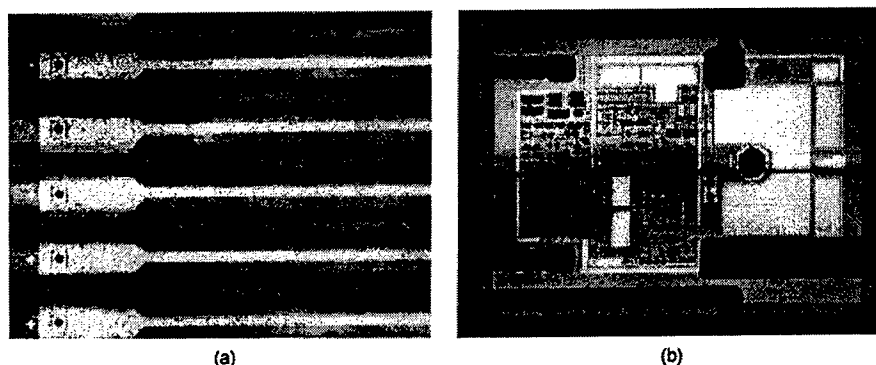


Figure 5. (a) Polymer waveguide array on VCSEL chips in a transmitter MCM and (b) a guide written by laser onto the detecting element (dark disk) of a high-speed receiver chip in a receiver MCM.

The POINT program (funded by DARPA and carried out by a consortium including GE, Honeywell, AlliedSignal, AMP, Columbia University, and UCSD) developed high-speed optical interconnects employing polymeric waveguides. In particular, the program demonstrated high-speed parallel optical data links between transmitting and receiving MCMs on a board, and from board to board through a backplane. One aspect of this program is direct *in-situ* adaptive writing of optical waveguides on these modules using the technology described here. Figure 5(a) depicts a section of a high-density array of 32 multimode waveguides directly laser-written onto VCSEL chips in a transmitter MCM, and Fig. 5(b) shows a laser-delineated waveguide accurately positioned on the detecting element of a high-speed receiver chip.

Light can be coupled from VCSELs into guides and from guides into detectors using 45° mirrors. A conventional approach would involve incorporating bulk optic prisms in the waveguiding circuitry to achieve the desired path bending. That approach lacks the accuracy and simplicity desired in a manufacturing environment. The POINT approach is completely lithographic and consists of shaping the facets of the waveguides to form reflectors. This result can be obtained using a variety of processing techniques. One such process is excimer laser ablation, which produces mirrors with a smooth finish that achieved reflection efficiencies as high as 80% without the use of reflective coatings. Figure 6(a) depicts an array of guides terminated with 45° mirrors and Fig. 6(b) offers a close-up view of the end facet in one such guide. Figure 6(c) shows a 45° mirror ablated in a waveguide directly on top of a detector. An alternate technique to excimer ablation of 45° mirrors is the direct laser termination of waveguides as they are laser delineated (the desired angle is achieved by a careful balance of the proper scanning speed, laser power, and material contrast additives). An array of laser-written waveguides directly terminated with 45° mirrors is shown in Fig. 6(d). The curvature of these mirrors can be controlled from convex, to straight,

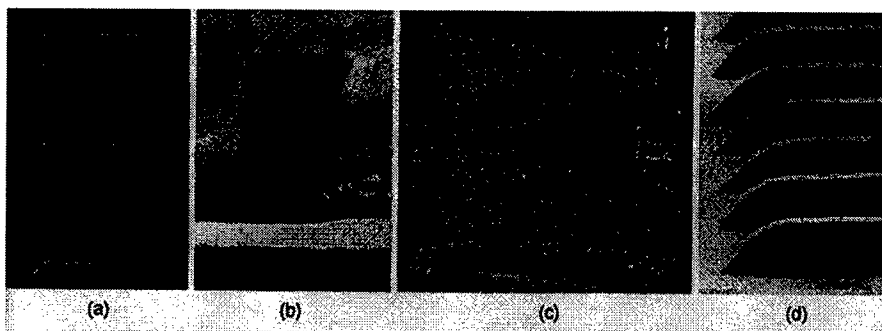


Figure 6. (a) An array of guides with excimer-laser-ablated 45° mirrors and (b) a close-up micrograph of one such mirror. (c) A 45° mirror ablated in a guide on top of a detector (dark disk). (d) An array of laser-written waveguides directly terminated with 45° mirrors.

to concave, while achieving any desired angle. The convex design is particularly useful since it helps focus the beam in addition to bending its path.

In the POINT program, the boards had waveguide array ribbons that communicated through a longer ribbon attached or laminated to a backplane. The board and backplane ribbons were terminated with MT-type connectors with tight dimensional specifications. In addition, the materials on both sides of the core (the cladding and the plastic substrate) had to be identical in order to avoid warpage. The micrograph in Fig. 7(a) shows the cross section of a ribbon segment, Fig. 7(b) depicts MT-connectorized waveguide array strips, and Fig. 7(c) shows the output of a connectorized strip when energized.

In one POINT transceiver demonstration, MT-connectorized transmitter and receiver modules were mounted on driver boards, and the two boards were mated forming a testable transceiver. Figure 8 shows such a link that includes AlliedSignal planar polymer waveguides, as well as micro-optical alignment

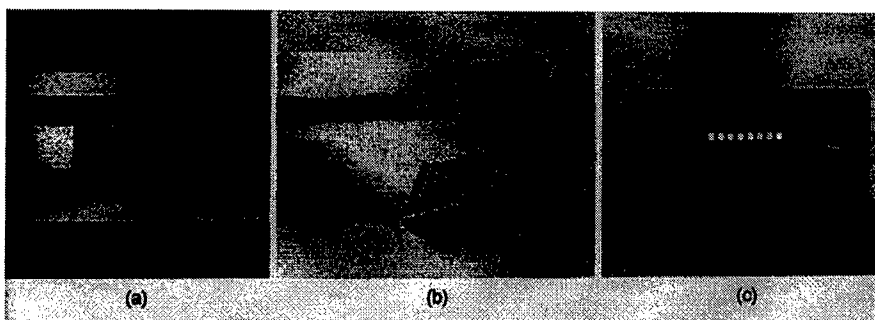


Figure 7. (a) Cross section of a waveguide array strip that is symmetric both geometrically and in terms of the materials above and below the core, (b) MT-connectorized polymer waveguide array strips, and (c) the output of one such strip when guiding.

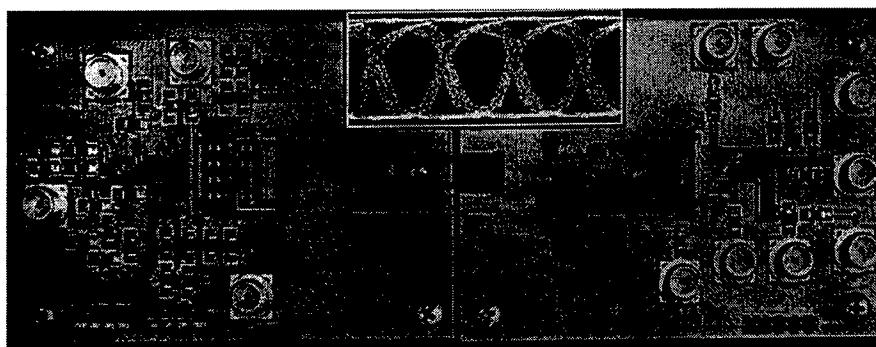


Figure 8. A POINT transmitter/receiver link that incorporates AlliedSignal polymeric components. The insert shows an eye diagram of a 1 Gb/s test performed on this transceiver.

elements which permit to passively align guide strips to the planar guides printed on the MCMs. High-speed tests were performed on POINT transceivers. An eye diagram with wide-open eyes at 1 Gb/s is depicted in the inset of Fig. 8. Open eyes were observed in tests run at up to 2.5 Gb/s. This development demonstrated, for the first time, an optoelectronic packaging technology for surface-emitting devices that leverages the planar fabrication and batch processing of the electronic IC industry for low cost and large volume manufacturing of optical interconnects.

Another natural role for low loss polymeric interconnects in computing cabinets is in backplane waveguides where dimensions can be reasonably large (say, 10-40 inches) and where the AlliedSignal state-of-the-art polymer loss values make it possible to produce the desired long waveguides while keeping the total insertion loss low. Several arrays of 200 9-inch-long waveguides were fabricated and they demonstrated excellent uniformity. In the final POINT demonstration, four such arrays were joined together to form a one-yard-long array of 200 guides. These flexible waveguide strips were attached to backplanes while allowing 90° bending to accommodate connection of daughter boards.

In a recent effort to mass-produce polymeric optical circuits at low cost, AlliedSignal developed a printing technique which, similarly to the newspaper industry, stamps the desired pattern on rolls of material. The substrate material of choice is a plastic (e.g., PET, polyimide), and printing was performed at speeds of 70 feet per minute. This process appears well-suited for high-volume production.

6. Conclusion

Optical interconnects offer a viable solution to the interconnect bottleneck problem facing the computer industry. Economic considerations make optical polymers the material system of choice for this application. Although no all-optical computers

are on the horizon, hybrid computers, in which nonlinear and logic operations are done electronically and massively parallel interconnections are done optically should be available commercially in the near future.

7. Acknowledgments

The author would like to acknowledge all the POINT team members for their contributions and DARPA for supporting the POINT program.

References

1. M. R. Feldman, S. C. Esener, C. C. Guest, and S. H. Lee, "Comparison between optical and electrical interconnects based on power and speed considerations," *Appl. Opt.* **27**, 1742 (1988).
2. L. D. Hutcheson, P. Haugen, and A. Husain, "Optical interconnects replace hardwire," *IEEE Spectrum* **24**, 30 (1987).
3. H. H. Arsenault and Y. Sheng, "Optics in computers," *SPIE TT8*, 54 (1992).
4. Y. S. Liu, "Lighting the way in computer design," *IEEE Circuits Dev.* **14**, 23 (1998).
5. L. Eldada, A. Nahata, and J. T. Yardley, "Robust photopolymers for MCM, board, and backplane optical interconnects," *Proc. SPIE* **3288**, 175 (1998).
6. L. Eldada and J. T. Yardley, "Integration of polymeric micro-optical elements with planar waveguiding circuits," *Proc. SPIE* **3289**, 122 (1998).

Development of RF equivalent circuit models from physics-based device models

S. Luryi

Dept. of Electrical and Computer Engineering, SUNY at Stony Brook, Stony Brook, NY 11794-2350, U.S.A

1. Introduction

Conventional RF modeling of device behavior is done by fitting parameters of a pre-conceived model of a given device (for example, the Ebers-Moll model of a bipolar transistor) to empirical RF data (for example, the measured scattering parameters). This approach fails in at least three situations:

- when one deals with a new type of device for which the device physicists have not done their homework;
- when miniaturization of standard devices brings about new physical phenomena not accounted for by the pre-conceived model; this situation is all too familiar and ranges from short-channel effects in a single MOS transistor to mutual interference between several closely spaced devices to effects of packaging and environment;
- when the device operation is stretched into a new regime, where the old pre-conceived model does not work; this situation includes high-power and/or high-frequency operation, or operation in an unusual physical environment, such as magnetic field, incident radiation, *etc.*

Modern device modeling codes enable the designer to simulate the behavior of nearly arbitrary three-dimensional semiconductor device structures with multiple electrodes. However, the output of such programs is not a lumped element model that can be used in the design of RF circuits. Rather, such programs produce a (hopefully) good simulation of an experimental situation, as described by the static and RF characteristics of the device. As far as RF modeling is concerned, the conventional use of device simulators boils down to imitating the RF scattering experiment from which one can obtain parameters of a pre-conceived lumped-element model. Clearly, this procedure suffers from the same limitations as fitting to experiment. This paper presents a different approach to this problem, based on physical device modeling without any pre-conceived circuit topology.

2. Physics-based device modeling

A good device modeling program solves semiconductor transport equations together with Poisson's equation and produces files that describe all the internal fields in the device — the electrostatic potential $V(x)$, the concentrations $n(x)$ and $p(x)$ of electrons and holes, the effective carrier temperatures $T_e(x)$ and $T_h(x)$, — as well as temporal variations of these fields. The solutions are usually discretized on a large grid, which in a modern simulator may involve millions of nodes. In a sense, such a grid may be viewed as an extremely complicated equivalent network. We should be able to extract workable equivalent circuit elements by a systematic reduction of the solution files — appropriately lumping together different nodes that are seen to be at the same potential, replacing streams of current by resistors, *etc.* Of course, this is precisely what a device physicist is doing implicitly while constructing an equivalent circuit based on a physical picture of the device. Our goal should be to *automate* this process! Ultimately, we need a robust procedure that would act as a postprocessor to device simulators — producing an equivalent circuit of any desired complexity to any multi-terminal semiconductor structure under any operating conditions, that the device simulator itself can model.

The range of validity of silicon device modeling is rapidly expanding to include new effects, such as those associated with heterostructure discontinuities and non-Boltzmann transport of hot-electrons. Also, device simulations have been successfully applied to electronic transport in compound semiconductor devices, and even to heterostructure lasers. Numerical accuracy of the simulation is not universally reliable in these unconventional regions of application, but qualitatively one often gets a good insight into the physics of device behavior.

In the approach proposed here, the relation between the physical picture provided by the simulator and that which exists in the real silicon structure is not to be challenged. Instead, we should take the simulator model as *infallible* and investigate means for reducing that model to an equivalent circuit.

While it is reasonably clear how to lump the simulator's grid nodes together into elements of capacitive, resistive, or inductive nature, the nature of current *generators* is less clear. Here, I am talking about *control generators*, which force a definite current through a branch of the lumped element circuit in response to a current (or voltage) condition at a different element of the circuit. For example, the Ebers-Moll model of a bipolar transistor contains the generator $G[I_E(I_B)]$ of emitter current controlled by the base current.¹ Similarly, an equivalent circuit of field-effect transistors usually contains a generator $G[I_{CH}(V_G)]$ of channel current controlled by the gate voltage.² How should such generators emerge during our simplification of the simulator's solution file?

One approach is to investigate nonlocal correlations between nodes. In particular, one can attempt to employ control theory methods and seek local negative differential impedance as a manifestation of one local region in the semiconductor device being controlled by another region. This latter approach is illustrated in Fig. 1, where a voltage source with some internal resistance r drives

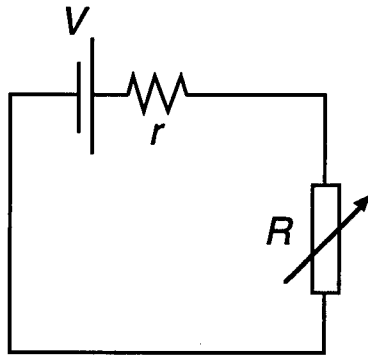


Figure 1. Let the externally controlled resistor R suffer a variation δR . Then the variation δV_R of the voltage drop on R will always be of opposite sign to the variation of current δI in the circuit. This negative "impedance" may serve as an indicator of an external control.

current through a variable resistor R (where R could represent any variable impedance device, such as a MOS transistor). When simulating this MOS transistor we can compute the variational "impedance" $\Re \equiv \delta V_{CH} / \delta I$ in response to (a) variation δV_G of the gate voltage and (b) variation δV_D of the drain voltage. As evident from the simple circuit in Fig. 1, in case (a) we find $\Re < 0$, which indicates that channel is controlled by the gate. By contrast, for case (b) we shall find $\Re > 0$, except for second-order effects associated with the small amount of control over the channel impedance by the drain voltage. It appears feasible to develop a robust procedure for identifying the generators in this way.

Another fundamental question to which we would like to provide an answer is when does the topology of an equivalent circuit change? Here, we do not mean incremental changes, such as a negligible capacitance at low frequencies becoming important at higher frequencies, but true topological discontinuous changes in the equivalent circuit, in response to a continuous variation of external parameters, such as the dc voltage, etc. We surmise that such changes will be associated with changes in the topology of equipotential surfaces and current networks in the full solution and recognizing these should involve some form of pattern recognition.

It seems that we should be able to obtain practical answers to such fundamental questions. Ultimately, we would like to develop a robust procedure that would act as a postprocessor to a comprehensive device simulator, producing an equivalent circuit of any desired complexity to any multi-terminal semiconductor structure under any operating conditions, that the device simulator itself can model.

I would like to emphasize again that this approach is radically different from the conventional ways of using a simulator to develop an equivalent circuit. What is done today is simply using the simulator as a convenient "experiment in the

bottle". Consider a three-terminal device, such as a transistor. Just as in an experiment, the simulation makes it possible to derive the scattering parameters and hence the impedance matrix Z_{ij} of the transistor. This matrix may be more convenient than that obtained from experimentally derived scattering parameters in that there is no question as to what is and what is not taken into account. That may simplify the guess of the equivalent circuit topology, but still the choice of topology requires an expert guess. Making the guess itself is in a sense external to the conventional RF model development process. Of course, if the guess is wrong, no fitting of lumped-element parameters will help. This difficulty serves as an indicator of the wrong circuit topology which can be used as a feedback to the expert, forcing him or her to improve the model. In contrast, we are mainly concerned with automating the choice of circuit topology. If the choice is made correctly, parameter validation should be quite routine.

3. Acknowledgments

This work is supported by a grant from the NSF/Industry Center for the Design of Digital and Analog Integrated Circuits (CDADIC).

References

1. P. M. Asbeck, "Bipolar transistors", in S. M. Sze, ed., *Modern Semiconductor Device Physics*, New York: Wiley Interscience, 1998.
2. Y. Tsvetkov, *Operation and Modeling of the MOS Transistor*, 2nd ed., New York: McGraw-Hill, 1999.

Contributors

Abramo, A.

DIEGM, Univ. di Udine, Udine, Italy

Akyüz, C. D.

Dept. of Physics, Brown University, Providence, RI, U.S.A.

Alieu, J.

Centre CNET — ST Microelectronics, Crolles, France

Anderson, W. T.

Naval Research Laboratory, Washington, DC, U.S.A.

Barabási, A.-L.

Dept. of Physics, Univ. of Notre Dame, Notre Dame, IN, U.S.A.

Bennett, A. J.

*Dept. of Electrical and Computer Engineering, Univ. of Toronto
Toronto, Ontario, Canada*

Bennett, H. S.

*Semiconductor Electronics Division
National Institute of Standards and Technology, Gaithersburg, MD, U.S.A*

Berggren, K.-F.

Dept. of Physics, Univ. of Linköping, Linköping, Sweden

Bois, D.

France Telecom, CNET, Meylan, France

Bouillon, P.

France Telecom, CNET, Meylan, France

Brémond, G.

INSA Lyon, UMR CNRS 511, Villeurbanne, France

Brown, D. E.

*Steacie Institute for Molecular Sciences, National Research Council of Canada
Ottawa, Ontario, Canada*

Bunyk, P.

Dept. of Physics and Astronomy, SUNY — Stony Brook, Stony Brook, NY, U.S.A.

Cahay, M.

Dept. of Electrical Engineering, Univ. of Cincinnati, Cincinnati, OH, U.S.A.

Calle, F.

Dpto. Ingeniería Electrónica, Univ. Politécnica de Madrid, Madrid, Spain

Calleja, E.

Dpto. Ingeniería Electrónica, Univ. Politécnica de Madrid, Madrid, Spain

Coonan, B.

National Microelectronics Research Center, Cork, Ireland

Crean, G. M.

National Microelectronics Research Center, Cork, Ireland

Cristoloveanu, S.

*Laboratoire de Physique des Composants à Semiconducteurs, ENSERG
Grenoble, France*

Datta, S.

Dept. of Electrical Engineering, Univ. of Cincinnati, Cincinnati, OH, U.S.A.

D'Avitaya, F. A.

CRMC2 — CNRS, Campus de Luminy, Marseille, France

Dehlinger, G.

Paul Scherrer Institute, Villigen-PSI, Switzerland

Derrien, J.

CRMC2 — CNRS, Campus de Luminy, Marseille, France

Dorojevets, M.

*Dept. of Electrical and Computer Engineering, SUNY — Stony Brook
Stony Brook, NY, U.S.A.*

Ekbote, S.

Dept. of Electrical Engineering, Univ. of Cincinnati, Cincinnati, OH, U.S.A.

Eldada, L.

Advanced Technologies, AlliedSignal Inc., Morristown, NJ, U.S.A.

Feinberg, E. A.

*W. A. Harriman School for Management and Policy, SUNY — Stony Brook
Stony Brook, NY, U.S.A*

Fiegna, C.

Dept. of Engineering, Univ. di Ferrara, Ferrara, Italy

Fjeldly, T. A.

*Center for Technology at Kjeller, Norwegian Univ. of Science and Technology
Kjeller, Norway*

Freund, L. B.

Div. of Engineering, Brown University, Providence, RI, U.S.A.

Furdyna, J. K.

Dept. of Physics, Univ. of Notre Dame, Notre Dame, IN, U.S.A.

Gennser, U.

Paul Scherrer Institute, Villigen-PSI, Switzerland

Goronkin, H.

Motorola Phoenix Corporate Research Laboratories, Tempe, AZ, U.S.A.

Grützmacher, D.

Paul Scherrer Institute, Villigen-PSI, Switzerland

Guillot, G.

INSA Lyon, UMR CNRS 511, Villeurbanne, France

Hartmann, R.

Paul Scherrer Institute, Villigen-PSI, Switzerland

Hess, K.

*Beckman Institute, Univ. of Illinois at Urbana-Champaign
Urbana, IL, U.S.A*

Hobler, G.

Institute of Solid State Electronics, TU Wien, Vienna, Austria

Holländer, B.

Institut für Schicht-und-Ionentechnik, Forschungszentrum Jülich, Jülich, Germany

Hong, Y. D.

Semiconductor R&D Center, Samsung Electronics, Yongin-City, Korea

Hu, Q.

*Dept. of Electrical Engineering and Computer Science
Massachusetts Institute of Technology, Cambridge, MA, U.S.A.*

Hwang, C. G.

Semiconductor R&D Center, Samsung Electronics, Yongin-City, Korea

Iñiguez, B.

*Electrical, Computer and System Engineering Dept.
Rensselaer Polytechnic Institute, Troy, NY, U.S.A.*

Iwai, H.

*Microelectronics Engineering Laboratory, Toshiba Corporation
Kawasaki, Japan*

Janowski, J. D.

Compaq Computer Corporation, Houston, TX, U.S.A

Jay, P. R.

Nortel Advanced Technology, Ottawa, Ontario, Canada

Jeong, M. Y.

*Dept. of Electronic and Electrical Engineering
Pohang Univ. of Science and Technology, Pohang, Kyungbuk, South Korea*

Jeong, Y. H.

*Dept. of Electronic and Electrical Engineering
Pohang Univ. of Science and Technology, Pohang, Kyungbuk, South Korea*

Johnson, H. T.

Div. of Engineering, Brown University, Providence, RI, U.S.A.

Johnson, S.

*Center for Solid State Electronics Research, Arizona State University
Tempe, AZ, U.S.A.*

Jorke, H.

Daimler-Benz Research Center Ulm, Ulm, Germany

Kasper, E.

Institut für Halbleitertechnik, Universität Stuttgart, Stuttgart, Germany

Kim, D. M.

*Dept. of Electronic and Electrical Engineering
Pohang Univ. of Science and Technology, Pohang, Kyungbuk, South Korea*

Kizilyalli, I. C.

Bell Laboratories, Lucent Technologies, Murray Hill, NJ, U.S.A.

Kosina, H.

Institute for Microelectronics, TU Wien, Vienna, Austria

Lam, C. F.

Electrical Engineering Dept., UCLA, Los Angeles, CA, U.S.A.

Lazzari, J.-L.

CRMC2 — CNRS, Campus de Luminy, Marseille, France

Lee, S. I.

Semiconductor R&D Center, Samsung Electronics, Yongin-City, Korea

Ledentsov, N. N.

A. F. Ioffe Physical-Technical Institute, St. Petersburg, Russia

Levner, D.

*Dept. of Electrical and Computer Engineering, Univ. of Toronto
Toronto, Ontario, Canada*

Li, J.

*Dept. of Electrical and Computer Engineering, Univ. of Toronto
Toronto, Ontario, Canada*

Likharev, K. K.

Dept. of Physics and Astronomy, SUNY — Stony Brook, Stony Brook, NY, U.S.A.

Lopinski, G. P.

*Steacie Institute for Molecular Sciences, National Research Council of Canada
Ottawa, Ontario, Canada*

Luryi, S.

Dept. of Electrical Engineering, SUNY — Stony Brook, Stony Brook, NY, U.S.A.

Luy, J.-F.

Daimler-Benz Research Center Ulm, Ulm, Germany

Lyding, J.

*Beckman Institute, Univ. of Illinois at Urbana-Champaign
Urbana, IL, U.S.A.*

Mantl, S.

Institut für Schicht-und-Ionentechnik, Forschungszentrum Jülich, Jülich, Germany

Mastrapasqua, M.

Bell Laboratories, Lucent Technologies, Murray Hill, NJ, U.S.A.

Matsuura, T.

*Research Institute of Electrical Communication, Tohoku University
Sendai, Japan*

Melloch, M. R.

*School of Electrical and Computer Engineering, Purdue University
West Lafayette, U.S.A.*

Merz, J. L.

Dept. of Electrical Engineering, Univ. of Notre Dame, Notre Dame, IN, U.S.A.

Moffatt, D. J.

*Steacie Institute for Molecular Sciences, National Research Council of Canada
Ottawa, Ontario, Canada*

Monroe, D.

Bell Laboratories, Lucent Technologies, Murray Hill, NJ, U.S.A.

Muñoz, E.

Dpto. Ingeniería Electrónica, Univ. Politécnica de Madrid, Madrid, Spain

Murota, J.

*Research Institute of Electrical Communication, Tohoku University
Sendai, Japan*

Naranjo, F. B.

Dpto. Ingeniería Electrónica, Univ. Politécnica de Madrid, Madrid, Spain

Naveh, Y.

Dept. of Physics and Astronomy, SUNY — Stony Brook, Stony Brook, NY, U.S.A.

Nurmikko, A. V.

Div. of Engineering, Brown University, Providence, RI, U.S.A.

Pacradouni, V.

*Dept. of Physics and Astronomy, Univ. of British Columbia
Vancouver, British Columbia, Canada*

Paddon, P.

*Dept. of Physics and Astronomy, Univ. of British Columbia
Vancouver, British Columbia, Canada*

Papadopoulos, C.

*Dept. of Electrical and Computer Engineering, Univ. of Toronto
Toronto, Ontario, Canada*

Park, Y. S.

Electronics Division, Office of Naval Research, Arlington, VA, U.S.A.

Patitsas, S. N.

*Steacie Institute for Molecular Sciences, National Research Council of Canada
Ottawa, Ontario, Canada*

Paul, D. J.

Cavendish Laboratory, Cambridge University, Cambridge, U.K.

Rakitin, A.

*Dept. of Electrical and Computer Engineering, Univ. of Toronto
Toronto, Ontario, Canada*

Razeghi, M.

*Dept. of Electrical and Computer Engineering, Northwestern University
Evanston, IL, U.S.A.*

Redmond, G.

National Microelectronics Research Center, Cork, Ireland

Reed, M.

Dept. of Electrical Engineering, Yale University, New Haven, CT, U.S.A.

Register, L. F.

*Beckman Institute, Univ. of Illinois at Urbana-Champaign
Urbana, IL, U.S.A.*

Regolini, J. L.

France Telecom, CNET, Meylan, France

Reitemann, G.

Institut für Halbleitertechnik, Universität Stuttgart, Stuttgart, Germany

Roenker, K.

Dept. of Electrical Engineering, Univ. of Cincinnati, Cincinnati, OH, U.S.A.

Sakuraba, M.

*Research Institute of Electrical Communication, Tohoku University
Sendai, Japan*

Sánchez, F. J.

Dpto. Ingeniería Electrónica, Univ. Politécnica de Madrid, Madrid, Spain

Sánchez-García, M. A.

Dpto. Ingeniería Electrónica, Univ. Politécnica de Madrid, Madrid, Spain

Sangiorgi, E.

DIEGM, Univ. of Udine, Udine, Italy

Selberherr, S.

Institute for Microelectronics, TU Wien, Vienna, Austria

Shur, M. S.

*Electrical, Computer and System Engineering Dept.
Rensselaer Polytechnic Institute, Troy, NY, U.S.A.*

Sigg, H.

Paul Scherrer Institute, Villigen-PSI, Switzerland

Skotnicki, T.

France Telecom, CNET, Meylan, France

Smith, T.

Compaq Computer Corporation, Houston, TX, U.S.A

Snowden, C.

Institute of Microwaves and Photonics, Univ. of Leeds, Leeds, U.K.

Souifi, A.

INSA Lyon, UMR CNRS 511, Villeurbanne, France

Syphers, D. A.

Dept. of Physics, Bowdoin College, Brunswick, ME, U.S.A.

Sze, S. M.

National Nano Device Laboratories, Hsinchu, Taiwan, R.O.C.

Tiedje, T.

*Dept. of Physics and Astronomy, Univ. of British Columbia
Vancouver, British Columbia, Canada*

Tuttle, B.

*Beckman Institute, Univ. of Illinois at Urbana-Champaign
Urbana, IL, U.S.A*

Unno, Y.

*Microelectronics Engineering Laboratory, Toshiba Corporation
Kawasaki, Japan*

Van Atta, R.

*Strategy, Forces, and Resources Division
Institute for Defense Analyses, Alexandria, VA, U.S.A.*

Wang, L.

*Electrical, Computer and System Engineering Dept.
Rensselaer Polytechnic Institute, Troy, NY, U.S.A.*

Wasshuber, C. L.

Texas Instruments Inc., Dallas, TX, U.S.A.

Watanabe, H.

R&D Group, NEC Corporation, Kawasaki, Japan

Weiner, D. D. M.

*Steacie Institute for Molecular Sciences, National Research Council of Canada
Ottawa, Ontario, Canada*

Weller, J.

Daimler-Benz Research Center Ulm, Ulm, Germany

Williams, R. S.

*Quantum Structures Research Initiative, Hewlett-Packard Laboratories,
Palo Alto, CA, U.S.A.*

Wolkow, R. A.

*Steacie Institute for Molecular Sciences, National Research Council of Canada
Ottawa, Ontario, Canada*

Xu, B.

*Dept. of Electrical Engineering and Computer Science
Massachusetts Institute of Technology, Cambridge, MA, U.S.A*

Xu, J. M.

*Dept. of Electrical and Computer Engineering, Univ. of Toronto
Toronto, Ontario, Canada*

Xu, Z.

*Electrical, Computer and System Engineering Dept.
Rensselaer Polytechnic Institute, Troy, NY, U.S.A.*

Yablonovitch, E.

Electrical Engineering Dept., UCLA, Los Angeles, CA, U.S.A.

Yoder, M. N.

Electronics Division, Office of Naval Research, Arlington, VA, U.S.A.

Young, J. F.

*Dept. of Physics and Astronomy, Univ. of British Columbia
Vancouver, British Columbia, Canada*

Zaslavsky, A.

Div. of Engineering, Brown University, Providence, RI, U.S.A.

Zinoviev, D.

Dept. of Physics and Astronomy, SUNY — Stony Brook, Stony Brook, NY, U.S.A.

Zozoulenko, I.

Dept. of Physics, Univ. of Linköping, Linköping, Sweden

Index

- A/D converters, 367
- adhesion, 18
- adsorption, 80-89, 273, 277-281, 283-284, 287
- advanced multifunctional rf system (AMRFS), 367
- AlGaAs, 165-166, 172, 235, 274, 335, 372, 383, 425-426, 431-432
- AlGaN, 385-386, 388-389, 391, 393-394, 397-398, 400-401, 406, 443, 446-450
- aluminum, 16-18, 21, 252, 254, 261-262, 321, 327, 335-336, 385, 393
- amorphous silicon, 213, 219
- amplifier, 126, 155, 164-166, 324, 329, 330-332, 336, 365-366, 374, 377, 449-450
- antimony, 140, 356, 358
- antireflection coating, 232
- arsenic, 151, 356
- asynchronous time mode (ATM), 5, 455
- atomic force microscopy (AFM), 237-239, 243-244, 247, 252-253, 266, 268, 398-399
- avalanche region, 294
- ballistic transport, 173, 250-251
- band structure, 155, 254, 431-432
- bandgap engineering, 35-36, 38, 72, 140, 155, 266, 361, 386, 443
- bandwidth, 9, 44-45, 56, 58, 196, 202, 259, 332, 367, 376, 407, 413, 416, 418, 452-453
- barium strontium titanate (BST), 16-17
- base current, 464
- base transit time, 170, 177
- base transport factor, 174, 176-177, 182
- BiCMOS, 62, 127-128, 184, 360, 372, 378
- bipolar junction transistor (BJT). *See* transistor
- bit error rate (BER), 197, 408, 417-418, 420
- Boltzmann constant, 308
- Boltzmann transport equation, 101, 251
- breakdown voltage, 390, 446
- buffer layer, 133-134, 144, 385, 393, 398-403, 406, 445-446
- cadmium, 364
- capacitance, 6, 16-17, 23, 26, 55, 61, 106, 115, 121, 127, 145-149, 174, 176, 208, 258, 298, 306, 308, 311, 314, 324, 332, 336, 359, 365, 417, 452, 465
- carbon, 141, 253-254, 262, 269, 279
 - C₆₀, 269, 274
 - nanotube, 253-254, 262, 269
- carrier lifetimes, 388
- CD-ROM, 10, 41, 346
- cell library, 4
- cellular automata, 244-248, 251, 255-256
- channel conductance, 92
- channel potential, 115
- channeling, 355-356, 358, 360
- charge injection transistor (CHINT). *See* transistor
- chemical beam epitaxy (CBE), 171, 180
- chemical-mechanical polishing (CMP), 3, 32, 35
- chip size, 3, 13, 23, 33, 35, 37-38, 61, 374, 452
- clock frequency, 4, 6, 13, 16, 22, 24, 34, 193, 197-199, 205, 333, 452
- CMOS, 4-6, 15, 21, 24, 31-32, 35, 48, 50-51, 55, 56, 59-64, 93, 101, 103, 106, 109, 112, 114, 124-128, 131, 133-134, 143, 150-151, 153, 183-190, 193, 195, 197, 200, 219, 245, 333-334, 348, 353-354, 365, 370-372, 375-377, 423, 453, 455
- code division multiple access (CDMA), 361, 407-410, 413, 416, 420-421
- coherent transistor. *See* transistor
- coherent transport, 176
- collector current, 167-168, 173-176

- computer, 8, 10, 19, 29, 34, 38, 41-47, 56, 58, 77, 103, 191, 193-197, 204-206, 213, 244, 249, 251, 256-259, 261, 265, 300, 302, 313, 333, 337, 339, 372, 374, 381, 423, 432-433, 451-453, 455, 457, 460, 461, 463
- conduction band, 133-134, 136-139, 175, 190, 223, 278, 325-327, 388, 433-434
 - offset, 133, 138-139, 326-327
- conductivity, 55, 176, 259, 266, 268-269, 272, 387-388, 401-403
- contact resistance, 166
- copper, 15, 17-18, 21, 27, 32, 204, 250, 273, 448, 453
- Coulomb blockade, 270, 274, 298, 307-308, 311, 314, 317, 320, 322, 336
- critical dimension (CD), 15-16, 18, 21-23, 28, 156, 173, 266, 451
- crosstalk, 325, 408, 415, 453
- cryomemory, 194, 195-196, 198-203
- current crowding, 159
- current drive, 19, 55, 61, 110, 115, 130, 185
- current gain, 166-168, 172-174
- current oscillations, 319, 330
- cutoff frequency, 127, 166-169, 171, 173, 174, 177, 185
- C-V measurement, 143, 147, 149, 151-152, 403

- deep-level transient spectroscopy (DLTS), 143, 145, 147, 151
- depassivation, 92-94, 96
- depletion, 16, 106-107, 109-111, 120, 216
 - charge, 106, 110, 120, 216
 - region, 106, 110, 216
- desorption, 83, 85, 86, 92-101, 279, 281-283, 287
- device area, 270
- diamond, 109, 284, 382, 443-444
- dielectric, 6, 15-18, 32, 37, 55, 236, 255, 324, 327, 385, 423-425, 428-431, 440, 444, 448
- diffusion coefficient, 208, 210, 402
- digital video disk (DVD), 41, 346, 384
- diode laser. *See* laser
- direct digital synthesizer (DDS), 363-366
- dislocations, 133, 364
- DNA, 56, 73
- domain, 70-73, 106, 112, 117, 340, 343, 407, 423, 434
- doping profile, 19, 133, 355, 357-358, 390
- double gate MOS, 106-107, 110, 113-124
- double heterostructure, 35, 232, 236, 389, 394
- drain current, 106, 111, 115, 145, 147, 150-151, 214, 215-216, 307, 309, 358
- drain-induced barrier lowering (DIBL), 110, 213, 216
- dual gate MOSFET. *See* transistor
- dynamic random access memory (DRAM), 10, 12-16, 18, 20-23, 29, 34-35, 38, 61, 65, 105, 114, 184, 194, 195-196, 289, 291-292, 296, 300-302, 319, 323-324, 328, 330-331, 333-334, 348, 353

- effective mass, 117, 134-135, 139, 144, 147, 150, 155, 161, 185, 224, 327, 435
- efficiency
 - injection, 170
 - power added, 166
- electric field, 96, 106, 117, 155, 185-186, 210, 215, 324-326, 328, 392, 429, 436, 442, 448
- electrical breakdown, 18
- electrically programmable read-only memory (EPROM), 294, 296, 303, 353
- electrically-erasable programmable read-only memory (EEPROM), 291, 294, 296, 303, 305, 312, 324
 - flash, 62, 291-292, 294-297, 299-300, 302, 303, 305, 306, 308, 311
- electromagnetic (E/M) systems, 363, 367
- electron temperature, 440
- emitter delay, 176
- environmental issues, 29
- epitaxy, 36, 38, 54, 79, 80, 83-86, 88-89, 105, 108, 128, 141, 144-145, 147, 153, 155-156, 163, 166, 167, 169-190, 221, 247, 326, 355-356, 360, 364, 368, 384-

- 385, 392-394, 397-398, 405-406, 423,
433, 435, 446-447, 450
etch-stop layer, 106
- fabrication technologies, 193, 195, 199, 234,
323, 330, 390
Fano resonance, 428-429
far infrared (FIR), 229, 233
feature size, 14-16, 48, 189, 250-252, 260,
323, 330, 333, 374, 391, 452
ferroelectric random access memory
(FRAM), 291-292
field-effect transistor (FET). *See* transistor
flash EEPROM. *See* electrically erasable
programmable read-only memory
(EEPROM)
flip-flop, 354, 365
foundry, 30, 50-54, 62
Fourier transform infrared (FTIR)
spectrometer, 437
Fowler-Nordheim tunneling. *See* tunneling
frequency domain, 408, 410-411
frequency modulation, 366-367
frequency multiplier, 433
frequency response, 91
- $\text{Ga}_{0.47}\text{In}_{0.53}\text{As}$, 182
 $\text{Ga}_{1-x}\text{In}_x\text{N}$, 386-389, 391
 GaAs , 37, 63, 72, 74, 77, 164-167, 183-185,
189, 224, 230, 232-237, 240, 245, 332,
335, 370, 372, 377, 382, 425-427, 429-
432, 434-435, 440, 442-444, 446-449, 453
 GaN , 165, 335, 346, 351, 363-364, 367-368,
381, 385-395, 397-406, 443-444, 446-450
gate capacitance, 16, 116, 121, 307, 311
gate delay, 4, 31
gate insulator, 218, 355
gate oxide, 15-16, 18-19, 31, 110, 116, 127,
143, 187, 328, 353-356, 360
gate stack, 16
gate structure, 116, 303, 305
germanium, 88, 132, 190, 192, 247
giant magnetoresistance (GMR), 347-348,
351
global positioning system (GPS), 41, 45
gold, 262, 268-270, 273-275, 321
graded base, 175, 177, 179, 182
grain boundaries, 216, 340, 364, 401
- Hall measurement, 401, 403, 405
hard disk (HD) storage, 56, 299-301, 332,
334, 340, 345
heavy hole, 135, 138-139, 155, 157-163,
169-170
heterointerface, 131, 145, 147, 152, 169
heterojunction, 36, 103, 139, 152, 155-156,
163, 165-166, 177, 184, 190-191, 365,
372, 377, 393, 443, 450
heterojunction bipolar transistor (HBT). *See*
transistor
heterostructure, 35, 72, 79, 88, 103, 125,
128-133, 138-139, 141, 152, 153, 155,
162, 164, 172, 182, 184, 186-188, -192,
227, 231, 234, 272-273, 339, 343, 348,
382-383, 386-388, 391, 393, 423, 425,
435, 442, 446, 464
 AlGaAs/GaAs , 165-166, 172, 274, 335
 InAlAs/InGaAs , 165-166, 171, 172
 InP/InGaAs , 165-166, 169, 171
 Si/SiGe , 124-125, 129, 132-134, 139,
140-142, 145, 147, 149, 152-153, 155-
158, 164, 175, 177, 182, 186, 190-192
 SiGe/Si , 129-130, 150, 163
heterostructure FET (HFET). *See* transistor
hexagonal, 252-253, 261, 341, 382, 401
high electron mobility transistor (HEMT).
See transistor
high-definition television (HDTV), 374
holographic storage, 345
hot carriers, 3, 31, 91-94, 96-97, 100-101,
106, 110, 129, 303
hot electron transistor. *See* transistor
hot-electron injection, 298
hybrid chip, 49, 51
hybrid technology multithreaded
architecture (HTMT), 194-196, 205-206,
333, 335, 337
hydrogen, 82-83, 89, 92, 98, 99-101, 108,
278, 281, 283, 285, 287

- ideality factor, 217
- III-V compounds, 33-38, 49, 54-55, 60, 63, 72-74, 109, 156, 183, 188, 190, 229, 233, 237-238, 245, 365, 387, 392, 424, 426
 - nitrides, 361, 381, 382-392, 397, 449-450
- II-VI compounds, 238, 245-246
- image processing, 8
- impact ionization, 110, 213, 215, 217
- implants, 19, 74, 188, 356-358, 360
- impurity diffusion, 79
- $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$, 169-170
- InGaAsP , 163-164
- injection angle, 180
- injection efficiency. *See* efficiency
- injection phase angle, 173, 176
- InP , 155, 163, 165-166, 169-172, 227, 235, 317, 322, 375
- input/output (I/O), 19, 125, 189, 197, 200, 359, 452
- integration, 1, 8, 16-17, 19, 22, 32-33, 35-38, 48, 54, 61-64, 68, 72, 74, 108, 125-126, 131, 143, 148, 151, 165, 183-184, 189, 199, 260, 265, 277, 289, 305, 316, 323-324, 330, 332, 348, 353-354, 360, 375, 397, 428, 451, 461
- interconnect, 3-6, 11, 15, 17-18, 21, 23, 26-27, 29-30, 32, 34, 48-49, 54, 61, 64, 68, 71-73, 75-76, 109, 126, 197, 204, 207, 210, 250, 265-266, 268, 324, 332, 451-456, 460-461
- interface states, 147, 188
- Internet, 9-10, 22, 29, 37-38, 41, 44-45, 50, 70, 393-394, 454
- interprocessor communication network, 194, 196
- intersubband transition, 433, 435, 436, 442
- inversion layer, 117-118, 122, 124, 153, 401, 406
- inversion layer mobility, 153
- ion implantation, 31, 145, 448
- ionized impurity scattering, 188
- islanding, 225-226
- Josephson junctions, 194, 197, 199, 202-203, 205
- Joule heating, 251
- kink effect, 115, 145, 214-219
- Langmuir-Blodgett films, 271, 275
- laser
 - ArF , 15, 21, 25
 - diode, 34, 235, 347, 351, 381, 384, 390, 391, 393, 395
 - KrF , 15, 21, 24
 - population inversion, 434, 435
 - printing, 384
 - semiconductor, 164, 223, 226, 346, 361, 377
 - stripe geometry structure, 160
 - vertical cavity surface emitting (VCSEL), 227-228, 230, 232-233, 347, 424, 455, 458
- latch-up, 106, 115
- lateral epitaxial overgrowth (LEO), 108, 355, 358, 360, 363-367, 381, 391-392, 447-448
- lattice constant, 23, 170, 224, 237, 384-385, 392, 424-425, 444
- lattice mismatch, 130, 140, 152, 155, 159-160, 226, 237, 239, 242, 364, 387, 391
- lattice temperature, 95, 223, 439
- layer-by-layer deposition, 80, 84, 86-87, 89, 239
- leakage current, 91, 106, 115, 149, 150-151, 216, 295, 314
- lifetime, 91-98, 100-101, 112, 229, 314, 341, 350, 374, 376, 389-391, 395, 428, 431, 448
- light hole, 136, 138, 140, 155, 158, 159, 160-162
- light-emitting diode (LED), 34, 233, 266, 383, 389-390, 395
- lightly doped drain (LDD), 145, 150
- lithography, 1, 3, 15, 21, 23-26, 29-32, 34, 49, 59, 68, 71-72, 126, 157, 244, 247, 250, 252-254, 257, 261, 265, 281, 317, 322-323, 328, 342, 350, 370, 375, 426, 451
- LO phonon. *See* phonons

- local oxidation of silicon (LOCOS), 3, 112
low-power electronics, 8, 103, 106, 111, 165, 172, 189, 302, 331, 349
low-pressure chemical vapor deposition (LPCVD), 79-80, 88-89
Luttinger liquid, 255
- Mach-Zehnder interferometer (MZI), 410-411, 413-414, 420
magnetic anisotropy, 341
magnetic force microscope, 340
magnetic random access memory (MRAM), 291-292, 339-340, 347, 349
magnetization reversal, 343, 350
magneto-optical (MO) storage, 56, 61, 72, 163, 188, 192, 213, 232, 252, 340, 343, 344-347, 371, 375, 391, 397, 411, 463, 466
mass production, 13, 15, 109, 112, 252, 317, 323, 385
maximum frequency of oscillation, 167, 171
memory cell, 16, 199, 202, 289, 297-299, 302, 305-311, 313, 316, 320-321, 324, 328, 330, 336, 348, 351, 359
mesa, 157-158, 160, 238, 242, 243-244
metal oxide semiconductor (MOS), 4, 31, 35-36, 56, 61, 72, 91-94, 96-97, 99-101, 105, 111-112, 115-116, 118, 123, 127, 152-153, 188, 213, 232, 249, 252, 260, 294, 303, 312, 336, 353, 360, 370-371, 375, 391, 397, 443, 463, 465-466
metal oxide semiconductor FET (MOSFET). *See* transistor
metal-organic chemical vapor deposition (MOCVD), 385, 393-394, 450
metal-semiconductor FET (MESFET). *See* transistor
micro-electromechanical systems (MEMS), 63, 74, 363, 366-367, 374
microelectronics, 1, 3-4, 6, 8, 10-12, 19, 21, 30, 48-49, 53-54, 58, 67, 77, 105-106, 111, 125, 132, 143, 155, 183, 190, 192, 210, 221, 229, 240, 249, 250, 260, 262, 277, 313, 321, 322-323, 334, 339, 348-349, 370, 372-375, 451
microprocessor, 6, 12-13, 19, 22, 43, 47, 58, 62, 113, 125, 127, 193, 195, 200, 250, 252, 260, 291, 299, 301, 375
microstrip lines, 197, 203
mobility, 35, 73, 111, 115-117, 121-124, 127, 132-133, 140, 143-144, 147, 149-152, 166-167, 177, 185, 188, 191-192, 213-214, 216, 358, 386, 388, 444
modulation, 37, 109, 140, 141-142, 188, 192, 232, 326, 327, 344, 351, 363, 366, 367, 391, 408, 411, 420
modulation-doped FET (MODFET). *See* transistor
modules, 19, 36-37, 61-62, 126, 184, 203-204, 333, 363, 434-435, 437-438, 440, 445, 451, 453, 458-459
molecular beam epitaxy (MBE), 54, 132, 134, 140-141, 142, 166, 171, 191, 234, 240, 246, 317, 321-322, 335, 351, 385, 388, 397-398, 402, 405-406, 435, 450
molecular electronics, 221, 245, 265, 349
monolithic integration, 125-126, 166
monolithic microwave integrated circuit (MMIC), 443, 445, 447-449
Monte Carlo simulation, 96, 123-124, 243, 305, 308, 311, 314, 357, 360
multimedia, 19-20, 37, 303, 379, 455
multiple quantum well (MQW). *See* quantum well
multithreading, 196, 206, 333
multi-valued logic, 265
- nanoelectronics, 1, 221, 249-250, 289
negative resistance, 110, 180
neural network, 70, 251, 255-256, 258-260, 423
noise, 111, 127, 130, 186, 207-212, 256-258, 324, 330, 336, 340, 342, 349, 353, 371, 374, 376, 388-389, 408, 411, 416, 417-418, 420, 453
nonvolatile memory, 289, 291, 300, 302, 313, 319, 335
- ohmic contact, 365, 402, 403, 440, 447
optical fiber, 11, 346, 407, 421, 453, 457

- optical interconnect, 63, 361, 451-458, 460-461
- optoelectronic integrated circuit (OEIC), 371, 374-375
- package, 63, 313, 363
- parasitic capacitance, 18, 34, 115, 365
- parasitic effects, 4
- passivation, 92, 128, 187-188, 448
- peak-to-valley ratio (PVR), 160
- permittivity, 6, 127
- personal computer (PC), 8, 10, 34, 38, 42-45, 46, 58, 299, 301, 333, 340, 378, 389
- petaflops, 103, 193, 195, 197, 199, 204-206, 333, 337
- petaops, 103, 193
- phonons, 95, 97, 122, 149, 152, 209-210, 223, 228-229, 234, 439
 - LO, 229, 439
- phosphorus, 31
- photodetectors, 381-383, 388-389, 393-394, 412, 416, 419
- photodiodes, 164, 388, 417
- photolithography, 31, 333
- photoluminescence, 134-137, 139-142, 230-231, 247, 386-387, 398, 400-401, 403-406, 425
- photonic bandgap material (PBM), 424-425, 430, 432
- photons, 223, 229, 235, 344, 346, 349, 423-425, 431, 434, 436, 439-440
- piezoelectricity, 382
- pinch-off, 306
- plasma, 15, 32, 86-89, 188, 192, 321, 347, 398, 433-434, 440-441
- platinum, 273
- Poisson equation, 117, 119, 123, 211, 434
- polymer waveguide, 451, 458-459
- polysilicon, 16, 88, 108, 129, 187, 213, 216, 218-220, 303, 353-354, 360
- population inversion. *See* laser
- potential well, 224
- power added efficiency. *See* efficiency
- power dissipation, 1, 5-6, 11, 35, 37-39, 68, 75, 106, 193, 199, 203, 251, 334, 453
- power electronics, 383-384, 397
- power gain, 166-168
- power-delay product, 171
- programmable read-only memory (PROM), 291, 294
- propagation delay, 34
- pseudomorphic structures, 133, 136, 139, 141-142, 185, 443, 446, 449
- pyramids, 225, 240-242, 247, 317
- pyroelectricity, 386, 392
- quantized subbands, 175
- quantum cascade laser, 442
- quantum computing, 37, 39, 74, 349
- quantum confinement, 39, 164
- quantum dot (QD), 36-37, 39, 164, 221, 223-238, 243, 245-248, 252, 255, 261-262, 265, 268, 274, 314, 321-322, 343, 391, 395, 423
- quantum well (QW), 84, 133, 135-136, 139-142, 144, 147, 155, 157, 160-161, 163-164, 175-176, 180, 186-188, 190, 227-229, 232-234, 236, 298, 387, 389-390, 393, 395, 421, 433, 435, 436-437, 441-442
 - multiple (MQW), 134-135, 137, 139, 390-391, 434-437, 440-441
- quantum-effect device, 79, 103, 106, 188, 265, 371
- radiative recombination, 226, 230
- rapid single flux quantum (RFSQ) logic, 195, 333
- rapid thermal annealing (RTA), 31, 188
- rate equation, 75, 81
- recombination, 110, 167, 227, 230, 285, 391, 400, 406, 434
- rectifier, 272, 330
- refractive index, 231-232, 382, 387, 424, 429, 457
- refractory metal, 16
- relaxation time, 122, 210, 224, 229
- reliability, 18, 21, 33, 63, 74, 91, 101, 244, 251, 295-296, 303, 374, 376, 383, 390, 401, 405, 443, 447-448, 450, 453

- resistance, 3, 18, 24, 26, 34, 111, 174-176, 177-182, 208, 210, 250, 270, 308, 311, 315, 347, 391, 417, 437, 447, 452, 457, 464
- resolution, 15, 21, 24-25, 31, 104, 148, 187, 224, 281, 313, 316, 322, 340, 345-346, 363, 375, 377, 398, 451
- resonance phase amplification, 174-175, 177, 180
- resonant tunneling, 71, 157, 161-162, 164, 176, 180, 182, 190, 207, 211, 266, 274, 375, 435, 442
- diode (RTD), 162, 164, 180, 182, 189, 190, 192, 207, 265, 375
- sequential tunneling, 328
- response time, 67, 258
- rf performance, 446
- ripening, 225, 238-240, 245-246
- roadmap, 1, 6, 11, 22, 29-30, 37, 47, 55, 59-60, 64, 77, 127, 132, 141, 205, 259-260, 323, 335, 369-373, 375-379
- saturation, 55, 60, 70, 84-85, 87, 127, 185, 213-214, 216, 249
- regime, 127
- velocity, 185
- voltage, 214
- scaling, 16, 18-19, 31, 55-56, 58-62, 64, 77, 113, 115, 186-189, 199, 214, 218-219, 249, 265, 323-324, 328, 330, 333, 339, 346, 349, 365, 424, 451
- factor, 265
- scanning tunneling microscopy (STM), 92-97, 100, 230, 237, 240-244, 266, 268-269, 275, 277-279, 281-287
- Schottky, 36, 207, 211, 271, 388-389, 394, 403, 445, 448
- barrier, 271, 394, 403, 448
- Schrödinger equation, 117
- self-aligned structure, 3, 79, 88, 187, 191, 234, 265-266, 321, 353, 355-356
- self-assembled quantum dot (SAQD), 237, 239, 242, 244, 246-247, 350
- self-limiting, 80, 83-85, 89, 308
- self-organization, 10, 11, 73, 224-225, 234-235, 237, 238, 246-247, 251-254, 257, 259-262, 266, 268, 271, 273-274, 343, 350
- Semiconductor Industry Association (SIA), 6, 22, 29, 47, 59, 64, 68, 245, 260, 330, 334, 370
- semiconductor laser. *See* laser
- sensors, 73-74, 105, 107, 109, 247, 266, 340, 346, 349, 374-375
- sequential tunneling. *See* resonant tunneling
- series resistance, 110, 112, 359, 392
- sheet density, 118, 121, 122
- short channel effects, 55, 88, 106, 110, 112-113, 115, 117, 143, 149-151, 213-214, 216, 218-220, 463
- shot noise, 207-212, 408, 416-417, 420
- Si_3N_4 , 16, 89, 183, 326-328, 448
- SiGe, 8, 55, 79-82, 88, 103, 129, 130-134, 139-153, 155, 160-161, 163-164, 177, 183-192, 365
- SiGeC, 133-136, 138-141
- silicon, 4, 6-10, 12, 16, 18, 20, 23, 30, 33-34, 36, 47-51, 53-54, 64-65, 69, 77, 79, 88-89, 92-93, 97-101, 103, 105-106, 108-109, 112-117, 119-124, 126, 128-132, 141, 143, 155, 163, 165-166, 171, 177, 183, 190, 192, 203, 213, 219, 229-230, 236, 240, 244, 245, 249, 271, 277-278, 280-281, 285-286, 291-292, 312, 321, 323, 327-328, 335-336, 350, 355-356, 360, 363-365, 367, 370, 372, 375, 377-378, 382, 384, 393-394, 405, 425-426, 448-451, 457, 464
- silicon carbide (SiC), 109, 129, 134, 139, 140, 262, 363-364, 384-387, 397, 405, 443-450
- silicon-on-insulator (SOI), 4, 19, 24, 31, 103, 105-119, 123-124, 215, 217, 219, 328, 360
- partially depleted, 109, 113
- single gate MOS, 115-122
- single-electron transistor (SET). *See* transistor

- SiO₂, 6, 17, 32, 55, 88, 92, 97-99, 108-109, 111, 115, 144, 147, 150, 183, 187-189, 293, 321, 325-326, 330, 355-356, 448
- slab waveguide, 425, 427, 429-430
- software, 7, 12, 29, 35, 38, 45-47, 58, 62, 69, 195, 198, 249, 256, 313
- solar cell, 213
- space charge, 434
- spectrum splitting, 236
- speech recognition, 8, 46
- spin, 74, 136, 247, 255, 277, 343-344, 347-349, 351
- spontaneous emission, 229, 431-434, 439
- spreading resistance, 167
- static induction transistor (SIT). *See* transistor
- static random-access memory (SRAM), 105, 114, 189, 192, 194-196, 201-202, 289, 291-292, 296, 301-302, 328, 348, 353-355, 358, 359-360
- stop band, 424
- stored charge, 293-295, 319, 325, 330
- strain relaxation, 134, 141, 144-145, 157, 159-162, 164, 187, 192, 225
- strained layers, 130-145, 147, 149, 151-152, 155-164, 185-192, 224-225, 233-234, 237, 240, 242, 246-247, 253, 343, 364, 384-385, 389, 391, 400, 446, 448
- Stranski-Krastanow, 239, 246, 253
- submicron devices, 92, 163
- subthreshold regime, 213
- subthreshold slope, 106, 110-111
- superchips, 1, 19, 62
- superconductors, 165, 194-195, 197, 200, 202, 204-206, 333, 336
- superlattice, 35-36, 84, 141, 164, 192, 234, 236, 424, 431, 441-442
- switching, 10, 23, 30, 38, 156, 194, 202, 245, 251, 324, 333, 335, 343-344, 350, 359, 361, 451-452, 455
- system integration, 125-127, 131
- system-on-a-chip (SOC), 19-20, 31, 35-36, 50, 64, 184, 205-206, 219, 235, 261, 273, 274, 285-286, 305, 312, 337, 392-393, 431-432, 450
- TaO₅, 16
- temperature
- growth, 238, 242, 364, 386, 398, 401-403
 - high, 105, 140, 145, 163, 188, 204, 223, 225, 356, 381-383, 397, 446, 449-450
 - low, 10, 31, 79-80, 83, 88-89, 133, 135, 139-140, 142-143, 149, 188-190, 195, 210-211, 223, 245-246, 282, 311, 330, 332-333, 388, 398, 400-401, 403-404, 405, 439-440
 - room, 30, 99, 106, 147, 175, 177, 189-190, 194, 196, 202, 204, 224, 227, 230, 235, 238, 245-246, 262, 265, 271, 274-275, 283-284, 298, 303, 311-312, 316-317, 320-321, 324, 326, 330, 332, 335-336, 350, 386-388, 391, 394-395, 397, 401, 435
- thermal budget, 128, 130, 133, 140, 143, 144, 147, 187-188, 356
- thermal conductivity, 111, 382, 397, 444, 447
- thermal noise, 23, 209-211, 417
- thermal resistance, 214-215
- thermal stability, 18, 188, 192, 457
- thermionic emission, 167, 169, 190, 326
- thin film transistor (TFT). *See* transistor
- threshold, 18, 53, 76, 91-92, 94-95, 106, 110-111, 114, 116, 119-121, 124, 127, 148-149, 152-153, 163, 169, 214, 216, 227-229, 234-236, 282, 293, 297, 305-308, 324, 358-359, 382, 390, 408, 427
- voltage, 92, 94, 106, 110-111, 114, 116, 119-121, 124, 127, 148-149, 152-153, 214, 216, 293, 297, 305-308, 358
- time domain, 408-409, 411
- time-dependent dielectric breakdown (TDDB), 295
- transconductance, 109-110, 115, 121, 127, 141, 185-186, 191, 216, 329, 445-446
- transfer characteristics, 307, 358
- transistor, 3, 6, 8, 10-13, 19, 23-24, 31, 33, 34-35, 37, 42, 47, 49, 55, 58, 61, 62, 64, 67-68, 71-72, 77, 91, 93, 98-101, 105-107, 109-114, 124-125, 127, 129, 131, 141, 143-152, 155, 165-166, 169-170,

- 173-177, 179, 181-186, 188, 191-193,
196, 213, 218-220, 245, 249, 250, 260-
261, 266, 272, 277, 285, 296, 298, 305,
311-312, 318-319, 323, 325, 328-330,
335-336, 351, 353-355, 358-360, 365,
374, 377, 433, 443, 445, 451-452, 454,
463, 465
bipolar junction (BJT), 111, 113, 127,
129, 165, 173-174, 182, 190, 301, 360,
365, 463-464, 466
charge injection (CHINT), 125, 129-130
coherent, 176, 182
field-effect
 heterostructure (HFET), 186-188, 190,
 192
 metal oxide semiconductor (MOSFET),
 16, 18-19, 23-24, 31, 37, 79, 88, 91,
 93, 96, 97, 100-101, 105-107, 109-
 115, 117-120, 123-124, 133, 139,
 141, 143, 152-153, 186, 190, 213,
 215, 217, 219, 260, 294-295, 306,
 328-330, 355, 359, 443, 446
 dual gate, 19, 123
 metal-semiconductor (MESFET), 190,
 374, 443, 445-446, 448-450
 GaAs, 37, 184-185
 modulation-doped (MODFET), 133,
 140-141, 185-186, 188-191
field-effect (FET), 71, 93, 124, 163, 184-
185, 190-191, 214, 216, 328-332, 336,
371, 374, 443, 445, 449-450, 464
heterojunction bipolar (HBT), 103, 125,
129, 130-131, 133, 141, 155, 156, 159,
160, 163, 165-167, 169-172, 177-178,
180, 182, 184, 190-191, 365, 372, 374-
375, 377, 443, 447, 450
 AlGaAs/GaAs, 166, 172
high electron mobility (HEMT), 171, 184-
185, 375, 443, 446-450
hot electron, 171
single-electron (SET), 6, 55, 212, 265,
305-312, 316, 329-332, 335, 336
static induction (SIT), 443, 445, 448-450
thin film (TFT), 213-220
transit angle, 173
transit time, 173-175, 371, 433
trench, 3, 16, 106, 129, 188, 324
tungsten, 17
tunnel diode, 188
tunnel junction, 212, 269, 274, 305-307,
311-312, 314, 317, 321, 335, 336
tunneling, 77, 156, 158, 160-162, 166, 169,
172, 189, 207, 210, 212, 214-216, 251,
265-266, 269, 274-275, 278, 282, 292,
298, 308, 311, 314, 321, 325, 328, 335-
336, 348, 351, 447
 Fowler-Nordheim, 216, 292, 294-296,
 298, 306, 325, 328
two-dimensional electron gas (2DEG), 210-
211, 386, 393, 446, 448
unified charge control model, 214
unilateral gain, 177, 180, 182
valence band, 133-134, 136-139, 143-144,
147-149, 152, 155-156, 169, 185, 190,
223, 386, 402
 offset, 138-139, 143, 147-149, 152, 386
vapor-phase epitaxy (VPE), 385, 392-393
vertical cavity surface emitting laser
 (VCSEL). *See* laser
very large-scale integration (VLSI), 31-32,
47, 49, 51, 53, 103, 114, 123, 152, 199,
206, 240, 244, 250-252, 260, 265, 303,
323-324, 333, 336, 360, 378, 452
wafer size, 1, 3, 15, 33, 68, 372, 384
wavefunction, 37, 99, 117, 122-123, 236,
251, 349
wavelength division multiplexing (WDM),
407, 411
wide bandgap semiconductor, 363, 384, 386,
449
wire pitch, 6, 16
x-ray diffraction (XRD), 134, 398, 400-401
ZnSe, 231, 236-238, 240, 245, 246-247